Rationality and Time

A Multiple-Self Model of Personal Identity over Time for Decision and Game Theory

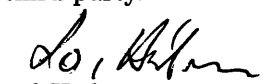
Conrad Heilmann

A thesis submitted to the Department of Philosophy, Logic and Scientific Method of the London School of Economics and Political Science for the degree of Doctor of Philosophy, December 2010.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it). The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without the prior written consent of the author. I warrant that this authorization does not, to the best of my belief, infringe the rights of any third party.



Conrad Heilmann

Abstract

This thesis presents extensions to formal theories of rationality in order to analyse intertemporal decisions. It offers multiple-self models of the decision-maker's personal identity over time. These models complement decision and game theory and are used to develop the new accounts of time discounting, backward induction, and preference change that are presented in this thesis.

The first part of the thesis develops multiple-self models of personal identity over time. These models depict a rational decision-maker as a series of different but interconnected temporal selves. The models allow one to relax the assumption that a rational decision-maker is a diachronically stable entity. Moreover, they structurally cohere with key problems and distinctions in theories of personal identity over time.

In the second part of the thesis, three problems of time in decision and game theory are analysed. Firstly, the problem of time discounting is considered. General foundations of time discounting are given in a measurement-theoretic framework. In the multiple-self interpretation of a decision-maker, the discounting factor represents the degree of connectedness between temporal selves in a person. Secondly, the reasoning method of backward induction in interactions over time is considered. Sufficient conditions for backward induction are given by formulating a belief revision policy on the basis of intrapersonal connectedness of players. Thirdly, preference change is considered. A new characterisation of diachronic inconsistency in terms of conflicts in intrapersonal connectedness is given.

The multiple-self models presented here allow one to represent the internal temporal structure of decision-makers. They capture problems of the interplay between rationality, identity, and time, thereby elucidating new accounts of time discounting, backward induction, and preference change. More generally, this thesis offers a new approach to modelling the intertemporal aggregation of value, which possesses broader relevance for decision theory, the foundations of economics, social epistemology as well as environmental ethics.

Overview

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to thank my supervisors, in particular Richard Bradley. I have benefited enormously from his patience, intellectual empathy, sharp (and quick) feedback as well as many astute suggestions. I am grateful for his outstanding support which went such a long way in helping me to develop and improve on this project. It was also a real pleasure to work with him on a practical and day-to-day basis. I would like to thank Katie Steele and Franz Dietrich for their co-supervision, generously offering their invaluable help and advice. I feel privileged to have worked with such a committed, encouraging, and kind team of supervisors.

I would like to thank my collaborators Christian W. Bach and Philip Cook for many hours of enjoyable, stimulating, and fruitful work. I am also indebted to Luc Bovens, Christian List, Mauro Rossi, and Olivier Roy, who offered so much help, advice, and encouragement all along.

Thanks to all members and seminar participants of the LSE Choice Group for creating a unique environment which has educated me a great deal. Most key ideas in this thesis have been aired in front of this most helpful, engaging, and critical of audiences and greatly improved as a result. The same holds for the Decisions, Games & Logic (DGL) workshops and the LSE PhD Seminar in Philosophy. I also want to thank organisers and participants of seminars and conferences for many helpful discussions at Amsterdam, Canberra, Eindhoven, Groningen, Norwich, Osnabrück, Paris, Prague, St. Andrews, and Sydney, where I have presented material contained in this thesis.

Thanks to everyone at the LSE Department of Philosophy, Logic and Scientific Method, the Centre for the Philosophy of Natural and Social Science (CPNSS), and the Philosophy Programme at the ANU in Canberra. I am also grateful for financial assistance from the German Academic Exchange Service (DAAD), the Academic Merit Foundation of German Businesses (sdw), and the LSE Research Studentship Scheme.

I have also very much benefited from discussions with and advice from Jason Alexander, Nicholas Baigent, Adam Brandenburger, Jake Chandler, Mark Colyvan, Tom Cunningham, Greg B. Davies, Boudewijn de Bruin, Marilena Di-Bucchianico, Foad Dizadji-Bahmani, Phillip Dorstewitz, Jacques Duparc, Damian Fennell, Marc Fleurbaey, Wulf Gaertner, Isabella Guerra, Brian Hill, James Joyce, Fabien Medvecky, Matteo Morganti, Ivan Moscati, Ittay Nissan, Alice

Obrecht, Eric Pacuit, Matt Parker, Andrés Perea, Jan-Willem Romeijn, Kai Spiekermann, Chris Thompson, Johan van Benthem, Martin van Hees, Alex Voorhoeve, Jon Williamson, and Stuart Yasgur.

Special thanks to Alice and Marilena for their congenial companionship in the PhD programme, and to Foad. Finally, sincere thanks to my parents, my sister and my brother, and so many friends and family that it is impossible to list them all. You know who you are!

Above all, thank you, Sarah.

Introduction

Chapter 1

Introduction

1.1 Intertemporal Decisions

Time plays an important role in many decisions. Indeed, intertemporality is prominent when deciding between small gains in the short-term and large gains in the long-term, and when deciding whether to do something earlier or later. Such intertemporal decisions play a highly significant role in the everyday life of individuals as well as in collective decision-making. For individuals, decisions about education, career path, migration, or investment in housing come to mind. For collectives, such intertemporal decisions include investments in education, pension systems and infrastructure, as well as dealing with environmental problems like climate change or biodiversity. Intertemporal decisions are important because they have a profound influence on the lives of the individuals and collectives concerned with them.

Consider the collective decision of how to deal with climate change. This decision has a strong intertemporal aspect, as the available courses of action differ in how the costs and benefits associated with them are distributed over time. One possible course of action is to incur costs in the short-term, by adopting measures to reduce carbon emissions, such as high taxes on fuels and investing in technology such as renewable energy. Such costs in the short-term might be outweighed by future benefits; for instance, if natural disasters and other possible side effects of global warming are mitigated. Other courses of action are also possible, such as not incurring costs in the short-term by not adopting any measures to counter climate change, with possible higher costs in the long-term. Assessing those different courses of action requires us, amongst other things, to

specify in what sense future costs and benefits of decisions are significant today. Do future benefits count less than present benefits, or are they to be evaluated on equal terms? How are we to factor in the possibility that what we think beneficial or costly today might not be considered so in the future?

Similar questions arise in intertemporal decisions of individuals. Consider an undergraduate student who decides whether to pursue further study or get a job instead. The consequences associated with each of these prospects will occur at different times. Moreover, it is likely that the two prospects have quite different consequences in the short-term and in the long-term: when pursuing further study, the student might have less money in the near future, and possibly a loan to pay after her studies. But in the long-term, the student might be able to earn more money because of her higher qualifications and she might lead a more satisfied life because of that. By getting a job instead of further study, she will earn more money in the short-term, but she might not be able to earn so much in the long-term. Even if the student is clear about how she values the kinds of jobs and studies available to her, the intertemporal aspect of the decision might still puzzle her. Should she think of the consequences in the far future as less valuable? How should she account for the fact that she might change her mind later about one of the possible courses of action? When individuals make decisions with a long-term impact, such aspects will matter a great deal.

The above examples suggest that intertemporal decisions are important and give rise to complex theoretical and practical questions. Not all of the questions that can be raised about such decisions will be treated in this thesis – indeed, I limit its scope to discussing three particularly interesting problems of intertemporality in decisions and games, and more generally to suggesting extensions to standard decision theories to better deal with such problems of intertemporality. Before discussing the latter, we turn to introducing the three particularly poignant problems of intertemporality in decisions and games.

1.2 Three Problems of Time in Decisions and Games

Intertemporal decisions with a much smaller significance than handling climate change or choosing a career path already exhibit three interesting problems of intertemporality. Take the decision of a group of friends whether to go out for dinner tomorrow night or rather next week. Even in such everyday decisions,

consequences occur at different times, which raises the question of how this fact impacts evaluations of decision-makers. More specifically, we can ask the following three questions.

(i) What is the significance of the fact that one dinner takes place later than the other one?

(ii) In what sense can decision-makers anticipate and deal with the fact that other decision-makers might surprise them over time?

(iii) How are decision-makers to deal with the fact that they might change their minds about possible courses of action?

These are the kinds of questions raised by intertemporality in decisions and games that this thesis addresses. In the following, we look at these three problems in more detail, and explain each of them by using a variant of the dinner example just introduced.

1.2.1 Temporal Distance and Time Discounting

The first problem of intertemporality is called the problem of *temporal distance.* Consider the dinner example. Suppose that the comparative quality of the dinners is not at issue; the two dinners are the same, except in when they will actually take place. What can we say about this temporal distance between the two dinners? Intuitively, there is an obvious difference between the two prospects of having dinner at different times. Hence, it is possible that the friends react in a different way. For instance, one of the friends could dislike waiting for social occasions and hence be rather impatient about the dinner taking place. In contrast, another one of the friends might get a lot of pleasure out of knowing that the dinner will be taking place, enjoying the anticipation of the occasion. Furthermore, some of the friends could foresee a lot of work-related commitments in the next week and think it unlikely that they will be able to join for the later dinner. Others amongst the friends could fear that not everyone will make it to tomorrow's dinner due to the short notice, and so on. That is to say, the mere fact that alternatives are different as to when they materialise can have a strong impact on their evaluation.

Many of the aforementioned phenomena associated with temporal distance can be accommodated straightforwardly. For instance, those dealing with uncertainties and those dealing with people's different subjective attitudes can be

modelled by standard decision-theoretic frameworks. Yet, once all such features are taken into account, it remains an open question as to whether temporal distance as such can change our goodness evaluations.

Here, the heavily contested concept of *time discounting* has been proposed to deal with temporal distance, by postulating weights on future outcomes to reduce their value (Frederick *et al.*, 2002). One question of this thesis is to develop foundations of time discounting, giving an exact description of how the notion of time discounting can be described and specifying on what kinds of assumptions it rests. In economics, time discounting is frequently employed to deal with temporal distance, such that goods occurring later in time are weighted less than those occurring earlier. In this context, the question arises in what way should goodness be weighted according to temporal distance. This has led to debates about the correct interpretation and method of time discounting. On the other hand, in philosophy, the position is by and large that temporal distance should not have an impact on the evaluation of an outcome (Sidgwick, 1907; Rawls, 1971; Broome, 1991, 1999), with Parfit (1984) being the most famous exception to this view. This thesis asks what kinds of evaluation of temporal distance can be represented by time discounting functions. In order to answer this question, we develop a general representational framework for time discounting that allows us to clarify existing theories, making transparent the requirements for the construction of well-founded time discounting functions.

1.2.2 Interaction over Time and Backward Induction

The second problem of intertemporality is called the problem of *interaction over time*. Suppose there is an element of strategic interaction between the friends who want to participate in the dinner: will everybody who has committed to coming turn up, and is one of the dinners more likely to attract a larger number of friends? For instance, it could be the case that more of the friends initially say that they prefer the later dinner but then do not come, as more attractive ways to spend the evening have become available to them in the interim. In deliberating about this problem and similar ones, each amongst the friends will make assumptions about the other friends' motivations. More specifically, hypothetical reasoning of this type will rely on what the friends know about each other and what kind of assumptions they make about each other. For instance, thinking about the other friends' previous actions in similar situations can become relevant when

considering one's own decisions. Furthermore, whether other decision-makers will be consistent over time and whether they will commit to decisions once made is also highly relevant to forming one's own evaluations. That is to say, hypothetical reasoning as described above requires one to entertain situations in which one is surprised by other decision-makers, such as when they do not stick to their plans.

The question of how to characterise decision-makers' interactive beliefs and possible changes in belief when facing surprises has been a key problem in models of hypothetical reasoning, in particular in discussions about the key reasoning method of *backward induction.* Epistemic game theory models dynamic interactions and provides a rich framework for its description with a focus on the role of knowledge (Brandenburger, 2007). Indeed, epistemic game theory characterises the epistemic assumptions of solution concepts in dynamic games, making transparent the hypothetical reasoning of the decision-makers. In this context, the reasoning method of backward induction is central, in which a decision-maker firstly entertains what she would choose at the last possible decision in a dynamic game and then works her way backwards to the beginning of the game (Perea, 2007). When determining possible courses of action in such a way, the decision-maker has to entertain situations which could only arise if another decision-maker is irrational (Stalnaker, 1998). How to keep the belief in the other decision-makers' rationality when entertaining such situations has been a pressing problem in the foundations of game theory (Binmore, 1987). This thesis shows how an enhanced representation of the temporal stability of decision-makers can improve the characterisation of backward inductive reasoning.

1.2.3 Temporal Dynamics and Preference Change

The third problem of intertemporality is called the problem of *temporal dynamics.* Let us focus on the deliberations of one decision-maker in the dinner example. Crucially, whatever decision one has taken, there will be time for reflection on the decision once the opportunities for the respective dinners arise and again when they have gone. It could be the case that after having decided to have the dinner next week, the decision-maker changes her mind about that decision. Furthermore, it could also be the case that she happens to lose interest in any social exchange with the friends, for instance because she realises that other friends are more important to her. It could also be the case that, upon learning that her two best friends can only come to either one of the dinners, she is

in a conflict as to which of the dinners she would prefer. In general, we can thus say that there are interesting temporal dynamics when decision-makers are given the chance to re-assess their decisions, revise their preferences, or remain in conflict about the decision. Some of those dynamics can be quite easily modelled within standard decision theory. In particular, decision theories allow for decision-makers to update their beliefs in light of new information. Yet, some of the dynamics mentioned before are not so easily analysable in terms of information, such as when decision-makers change their tastes. Such problems of changing and conflicting preferences remain contested in both philosophy and economics (Stigler and Becker (1977), Bradley (2009b)) .

A particular kind of preference change, commonly referred to as dynamic inconsistency, occurs when a decision-maker has contradictory preferences over time. One of the concerns of the thesis is to analyse *preference change* by an account that improves on the explanation of dynamically inconsistent preferences of decision-makers over time. In this context, Schelling (1980, 1984) and Ainslie (1992, 2001) have proposed to draw on the metaphor of persons as 'multiple-selves' in order to analyse dynamic inconsistency, such that different selves have opposing preferences. In the more recent behavioural economics literature, such approaches have been combined with hyperbolic discounting functions, modelling dynamic inconsistency as being produced by the interaction of short-sighted and far-sighted selves, for instance in Thaler and Shefrin (1981) and Fudenberg and Levine (2006). This thesis presents a modelling approach that is simpler yet achieves the same aims as the aforementioned ones, and makes transparent in what way the description of dynamically inconsistent decision-makers requires us to depart from normative decision-theoretic accounts.

1.3 Time in Decision and Game Theory

This thesis addresses the three problems of (i) temporal distance and time discounting, (ii) interaction over time and backward induction, and (iii) temporal dynamics and preference change. Considering these three problems also raises the question of how standard decision-theoretic accounts can be used to analyse them. This, in turn, leads to a general concern about how standard decision-theoretic accounts can be extended in order to better analyse the three problems of intertemporality.

1.3.1 Extensions for Decision Theory

Standard decision-theoretic accounts take a subjectivist approach to analysing decisions. They embark from a sparse and idealised representation of an individual's mental state by way of a pair of probability and value functions. The probability function represents the individual's beliefs and the value function represents the individual's desires. Decisions can then be analysed on the basis of such two-factor models. More specifically, the expected goodness of possible courses of actions can be evaluated by combining the beliefs of the decision-maker with the desirability she assigns to the different options. Formulating rationality conditions on the structure of beliefs and desires yields accounts of rational decision-making. Moreover, decision theories not only present powerful tools to analyse decisions, they also give subjectivist foundations for economics and social sciences. Indeed, the latter can be based on a decision-theoretic analysis of their choices, interactions, and collective actions.

Yet, standard decision-theoretic representations of individuals do not offer tools to explicitly consider separate attitudes towards the future, such as those implied by time discounting functions, and possible changes in tastes. Hence, it is natural to investigate how standard decision-theoretic frameworks can be enriched in order to widen their scope and applicability. Let us now recall the three problems of intertemporality and their discussion in the literature. It seems that all three problems have in common that a direct application of a standard decision-theoretic framework does not give conclusive answers to them.

Consider the problem of time discounting for temporal distance. Decision-theoretic representations permit a wide range of individual desirability attitudes: it is just as permissible to take the future as less important, as equally important, or even as more important than the present. That is to say, attitudes towards the future can simply be regarded as being a matter of personal taste. However, note that in analysing real-world intertemporal decisions, the question arises whether evaluations of temporal distance can motivate time discounting factors that lower goodness evaluations at future time points. Standard decision-theoretic accounts do not directly permit us to discuss the evaluation of time distance as a separate concern, contrary to how it is often regarded in real-world intertemporal decisions, such as climate change for collectives, or career path for individuals. This raises the question what kind of extensions to decision theory are required to discuss time discounting in its framework.

Now consider the problem of backward induction in interactions over time. Standardly, game theory assumes the sequential stability of the preferences as well as plans of action of individuals. Yet, precisely such stability assumptions need to be locally relaxed in order to account for surprise information in hypothetical reasoning: if someone faces another individual who deviates from her plan of action, then this new information needs to be accommodated within the existing beliefs of the individual. Here, epistemic game theory provides tools to model hypothetical reasoning of decision-makers about each other's actions. However, the modelling devices of epistemic game theory require us to specify in what way individuals revise their beliefs about each other. This raises the question of how epistemic models of dynamic games can be amended such as to accommodate surprise information that is key to backward inductive reasoning.

Finally, consider the problem of preference change in temporal dynamics. Many applications of decision-theoretic frameworks embark from the assumption that an agent's preferences are stable over time. An important tool to deal with some types of changes is the so-called 'updating' of beliefs in light of new information or evidence. That is, by learning new propositions, an individual can revise her beliefs so as to correctly represent her knowledge about the world. However, changes of preference can also occur in less laudable circumstances, for instance, when individuals change their tastes without apparent motivation and when they contradict themselves over time. From a decision-theoretic point of view, such changes in preference are irrational. Yet, real-world decision-makers change their tastes and contradict themselves quite persistently, such as when failing to adhere to a healthy diet. This raises the question of how such preference changes can be described in order to better understand how individuals persistently violate the assumption of dynamic consistency.

The discussion of the three problems of intertemporality in the context of decision and game theory suggests that they not only present important questions from a practical point of view. They also fall outside the scope of a direct application of standard decision theories. This establishes a second key concern of this thesis, over and above providing accounts of the three problems of intertemporality: namely to ask how decision-theoretic representations of individuals can be enriched in order to enable analysis of intertemporal decisions.

1.3.2 Multiple-Self Models of Personal Identity over Time

This thesis proposes the modelling device of 'multiple-self models of personal identity over time' to suitably extend decision theory in order to analyse the three problems of intertemporality from a decision-theoretic point of view.

Multiple-self models introduced in this thesis depict a decision-maker as a collection of temporal selves, and characterise their degree of connectedness. More precisely, temporal selves capture the idea that a decision-maker exists at different points in time, with possibly changing characteristics. The notion of connectedness describes the degree of stability between temporal selves, i.e. the degree to which temporal selves have similar characteristics. In a general sense, these multiple-self models will allow us to capture decision-makers as temporally extended persons, offering precise descriptions of their stability over time. This provides a structure to express what to take as relevant about the changing nature of decision-makers.

From a decision-theoretic point of view, the concepts of temporal selves and connectedness can be interpreted reductively. On such an interpretation, at each point in time, the temporal self is depicted as a rational decision-maker that is constrained by standard decision-theoretic consistency assumptions. The degree of connectedness describes in what sense the different temporal selves are similar to each other. In Chapter 2, we will discuss in detail how such a reductive multiple-self model relates to standard decision-theoretic representations, and we will give different variants of multiple-self models that can complement decision-theoretic representations of decision-makers.

Furthermore, the multiple self-models proposed here are also closely related to theories of personal identity over time. Such theories of personal identity over time offer accounts of how persons change and how we can understand such change (Noonan (1989), Kolak and Martin (1991), Shoemaker (2008), Olson (2008)). More specifically, theories of personal identity over time ask how one can describe a person at different times as qualitatively a somewhat different person yet numerically (or quantitatively) as still the same person. We discuss those accounts of personal identity over time and take them as possible sources of motivation for modelling the influence of time on decision-making. We will argue in Chapter 3 that there is a structural coherence between multiple-self models and key distinctions and questions in theories of personal identity over time. Hence, we show how it is possible to constrain multiple-self representations

of decision-makers with insights from theories of personal identity over time.

More generally, the multiple-self models introduced in this thesis can be interpreted as relaxing the assumption of temporal stability that is often made in applications of decision theories. This suggests that while the multiple-self models depict insights into decision-makers' personal identity over time, we are not required to endorse them as metaphysical views of the ontology of decision-makers. Rather, the models can be viewed as structures that offer a more permissive representation of the decision-maker's deliberations about time. That is to say, in order to apply the aforementioned models as extensions to formal theories of rationality, no additional assumption is made other than to accept them as a method of relaxing stability assumptions that are often made in their application. Applying such a description of temporally extended decision-makers to intertemporal decisions allows much richer interpretations of the new accounts of time discounting, backward induction, and preference change presented in this thesis.

1.4 Thesis Overview

The thesis is divided into two parts. Part I develops the multiple-self approach, showing how multiple-self models can enrich decision theory (Chapter 2), and how they structurally cohere with philosophical theories of personal identity over time (Chapter 3). Part II provides three accounts of temporal problems in decisions and games: foundations of time discounting (Chapter 4), sufficient conditions for backward induction (Chapter 5), and dynamically inconsistent preference change (Chapter 6). Chapter 7 offers conclusions.

1.4.1 Part I

Part I develops multiple-self models of personal identity over time. It is argued that such modelling devices both present extensions to decision theories and structurally cohere with key distinctions in theories of personal identity over time.

Chapter 2. Multiple-Self Models. Chapter 2 discusses how decision theories usually represent decision-makers and how such representations can be enriched in order to capture changes in decision-makers over time. We introduce the device of multiple-self models that allows us to conceive of a decision-maker as a collection of temporal selves and their connectedness. In

a reductive interpretation, temporal selves are sets of preferences and their connectedness is determined by similarity of preferences. In a non-reductive interpretation, temporal selves are seats of a broader range of psychological features, such as preferences, memory, emotions, etc. We show how such an extended representation of decision-makers can capture the influence of time on decision-making and how it relates to existing proposals of understanding individuals as 'multiple-selves'.

Chapter 3. Personal Identity over Time. Chapter 3 shows how multiple-self models structurally cohere with key questions and distinctions in theories of personal identity over time. We suggest that those theories can be viewed as competing answers to three questions: (i) instances, the question of what is significant for a person to exist at a given point in time – this can be captured by the notion of temporal selves in a multiple-self model, (ii) persistence, the question of what is significant for a person to exist over time – this can be captured by the notion of connectedness in a multiple-self model, and (iii) criteria, the question of what establishes instances and persistence of persons – this can be captured by an interpretation of temporal selves and connectedness in a multiple-self model. That is, given a sufficiently rich specification of a multiple-self model, it can be motivated and constrained by specific theories of personal identity over time.

Multiple-self models of personal identity over time will be used in the second part of the thesis to improve our understanding of three problems of time in decision and game theory. While many of those discussions do not directly hinge on endorsing multiple-self models as a premise, we will show how for each of those problems, they offer valuable modelling devices that give additional insights into how intertemporality can be analysed by decision and game theory.

1.4.2 Part II

Part II discusses three problems of intertemporality in decision and game theory.

Chapter 4. Foundations of Time Discounting. Chapter 4 investigates how time discounting functions analyse temporal distance in intertemporal decisions. We identify two goals that theories of time discounting may have: one, postulating a correct time discounting function, and two, offering an

accurate underlying conceptual motivation. We proceed by presenting a general representation framework for time discounting which outlines the requirements that well-founded time discounting functions have to fulfil. This general framework is used to analyse both existing accounts of time discounting, as well as Parfit's dictum of time discounting because of a weak connectedness to future selves. More generally, the requirements for time discounting theories developed here demonstrate that time discounting factors are restricted in the kinds of conceptions they can express.

Chapter 5. Backward Induction. Chapter 5 analyses the problem of interaction over time; in particular, the sequential structure of dynamic games with perfect information. A three-stage account is proposed, that specifies set-up, reasoning and play stages of dynamic games. Accordingly, we define a player as a set of agents corresponding to these three stages. Moreover, the notion of agent connectedness is introduced which measures the extent to which agents' choices are sequentially stable. A type-based epistemic model is augmented with agent connectedness and used to provide sufficient conditions for backward induction. Moreover, an existence result is obtained ensuring that these conditions are indeed possible. Our epistemic foundation for backward induction makes explicit that the epistemic independence assumption involved in backward induction reasoning is stronger than usually presumed. Furthermore, in the three stage-account, players can explicitly be understood as multiple-selves, which permits one to interpret low agent connectedness as stemming from imperfect connectedness between selves.[1]

Chapter 6. Preference Change. Chapter 6 analyses temporal dynamics and gives an account of dynamic inconsistency. Two families of approaches to dynamic inconsistency are identified: firstly, those that use hyperbolic discounting functions to describe dynamically inconsistent decision-makers as myopic, and secondly, those that postulate multi-selves models that capture different motivations and time horizons which can lead a decision-maker to (fail to) control himself in the face of temptation. In order to achieve a simpler characterisation of dynamic inconsistency, we reconsider both hyperbolic discounting and multi-selves models in the more general model

[1]This chapter is based on a joint paper (Bach and Heilmann, 2009) with Christian W. Bach (University of Maastricht, Netherlands) to which both authors contributed equally.

of connectedness in the multiple-self. A simple specification of this model can motivate hyperbolic discounting, and an extended version of it can be used to reformulate the multi-selves models, using a less complex structure that can be better motivated. Moreover, the latter allows us to distinguish between conflicts in connectedness and conflicts in goodness evaluation.

1.4.3 Outlook

This thesis is divided into two parts, with the latter part focusing on three specific problems in intertemporal decisions and games, and the former part developing modelling devices that facilitate the analysis. However, this does not entail that all the accounts, arguments, and formal results in those three accounts necessarily *depend* in any way on accepting the modelling devices as premises. Naturally, the three accounts differ in how far they draw on the multiple-self models in their analysis.

In light of this, there are two modes of reading this thesis. Firstly, there is a 'thin' reading which focuses on the three accounts of intertemporality in decisions and games and understands them as isolated solutions to specific intertemporal problems. On this reading, Chapter 4 provides general foundations of time discounting, Chapter 5 proposes new sufficient conditions for backward induction, and Chapter 6 analyses dynamic inconsistency. Indeed, the individual chapters in Part II are intended as, by and large, self-contained treatments of received problems in their respective areas of enquiry.

Secondly, there is a 'thick' reading of the thesis, which centres around the multiple-models of personal identity over time developed in the first part. On this reading, those models can be taken as a general proposal for modelling intertemporality in decisions and games, for which the three accounts offered here are applications, demonstrating the usefulness of the multiple-self models through the additional insight they allow us.

Part I

Multiple-Self Models of Personal Identity over Time

Chapter 2

Multiple-Self Models

Summary. This chapter discusses how decision theories usually represent decision-makers and how such representations can be enriched in order to capture changes in decision-makers over time. We introduce the device of multiple-self models that allows us to conceive of a decision-maker as a collection of temporal selves and their connectedness. In a reductive interpretation, temporal selves are sets of preferences and their connectedness is determined by similarity of preferences. In a non-reductive interpretation, temporal selves are seats of a broader range of psychological features, such as preferences, memory, emotions, etc. We show how such an enriched representation can capture the influence of time on decision-making and how it relates to existing proposals of understanding individuals as 'multiple-selves'.

2.1 Introduction

Decision-theoretic accounts provide representations of decision-makers' states of mind. Indeed, their models combine a formal characterisation of beliefs and desires to analyse the decision-making of individuals. Such two-factor models are widely applied, and also used to provide foundations in economics and social science, for instance by motivating utility functions in economics, and by providing foundations for methodological individualism. The widespread use of decision-theoretic accounts is due to their sparse and therefore highly flexible representation of decision-makers: only some assumptions on the structure of beliefs and desires are needed to formulate models of rational decision-making.

Naturally, such models can also be used to analyse intertemporal decisions.

Yet, as briefly alluded to in the introduction of this thesis, the sparse structure of standard decision-theoretic representations does not enable us to discuss how to conclusively evaluate all aspects of the problems of (i) temporal distance, (ii) interaction over time and (iii) temporal dynamics in greater detail. This chapter briefly introduces the main features of decision-theoretic representations of decision-makers and proposes an answer to the following question: how can decision-theoretic representations be enriched so as to be applicable to the three problems of time decisions and games?

We begin the chapter by reviewing the basic elements of standard decision-theoretic representations of decision-makers, focusing on the capacity of such accounts to analyse intertemporal decisions. Extensions to decision theory are introduced that have been proposed in the literature to deal with some aspects of such decisions. Yet, further extensions of decision-theoretic frameworks are required in order to answer the kinds of questions about intertemporal decisions posed in this thesis.

In a second step, we introduce accounts from a that rather diverse literature that have attempted to provide more enriched models of individuals by drawing on the metaphor of the 'multiple-self' (Elster, 1986). Such accounts view individuals as consisting of several different selves in order to analyse their conflicts of motivation. However, as pointed out by Frederick *et al.* (2002) when discussing the role of multiple-self accounts in the analysis of intertemporal decisions, 'most of these multiple-self models have not been formalized'.

This chapter introduces 'multiple-self models' that enrich standard decision-theoretic representations of decision-makers to improve on this deficiency. They represent the temporal dimension of prospects by a model of the decision-maker as a collection of interconnected temporal selves. Such models are used to describe the decision-maker's attitudes to the temporal aspect of prospects. That is, multiple-self models provide tools to widen the scope of decision theory to analyse intertemporal decisions in greater detail.

The chapter proceeds as follows. Section 2.2 discusses how decision-makers are represented in decision-theoretic accounts and shows that extensions are required to deal with the three problems of time identified in the introduction of this thesis. Section 2.3 reviews how the notion of the 'multiple-self' has been employed in the literature. Section 2.4 introduces multiple-self models that characterise a temporally extended decision-maker as a collection of temporal selves which

are connected to each other, and proposes different kinds of interpretation of connectedness. Section 2.5 briefly concludes.

2.2 Time and Decision Theory

This section reviews standard accounts of the representation of decision-makers and some proposed extensions, focusing on the capabilities of such accounts to analyse intertemporal decisions. In particular the problems of temporal distance, interaction over time, and temporal dynamics are considered. This sets the scene for presenting the device of multiple-self models later in this chapter.

2.2.1 Standard Decision-Theoretic Representations

Decision theories offer tools for characterising how decision-makers evaluate prospects. In general, prospects are assumed to be rich descriptions of 'complete world histories', i.e. sets of possible worlds. For example, take the prospect of going to the beach. A maximally rich description of this prospect encompasses all events that can possibly be associated with it, such as going swimming, finding a space on an overcrowded beach, eating ice-cream, seeking shelter from the rain, and so on. Decision theories formulate two-factor models that combine beliefs and desires in an attempt to characterise the attitudes of decision-makers to such prospects. Indeed, such two-factor models of decision-makers' states of minds are employed to give a quantified characterisation of an individual's beliefs and desires.

Imagine a decision-maker who evaluates the prospect of going to the beach. In a decision-theoretic analysis, her attitudes to these prospects are characterised by her beliefs and desires. That is, her beliefs, such as how likely it is that the beach will be overcrowded, are combined with her desires, such as how much the individual enjoys swimming. For instance, her desire to go swimming might be combined with a high degree of belief that the beach will be overcrowded. If other prospects are available to the individual, such as staying home, then those might be evaluated as better overall by her than going to the beach because of her high degree of belief that the beach will be overcrowded.

In a general sense, an individual's attitudes to prospects can be described by combining her beliefs and desires. More formally, decision-theoretic accounts postulate structural conditions on beliefs and desires sufficient for representing

the decision maker's beliefs by a probability function p, and her desires by a value function v. These functions are defined over a set of prospects, such that $\langle p, v\rangle$-models give a quantified representation of beliefs and desires with regards to prospects (Bradley, 2009a,b). The above ingredients of a decision-model can be understood as the core idea that underlies decision-theoretic frameworks. This core idea has been specified in greater detail in different decision theories.

In order to obtain such a representation, many decision theories embark from the notion of a preference ordering over a set of prospects. That is, decision-makers are assumed to rank different prospects in terms of their desirability. If such preference orderings satisfy some structural assumptions, such as weak ordering and independence conditions, they can be represented by a utility function. If the latter is weighted with the decision-maker's probability function, we obtain an expected utility representation of the decision-maker's preferences. That is, a decision-maker prefers prospect A over prospect B iff a larger expected utility is associated with prospect A than with prospect B. To obtain expected utility, von Neumann and Morgenstern (1944) employ an objective notion of probability, whereas more recent frameworks, such as Savage (1972), Jeffrey (1983), Joyce (1999), and Bradley (2007b), interpret the probability function as a representation of subjective degrees of belief. In the following, all those aforementioned decision theories will be referred to as $\langle p, v\rangle$-representations of a decision-maker's mental states. Such representations of belief and desire are, in their most basic forms, reductive and sparse as they only require a few structural assumptions. They offer a highly idealised and flexible representation of the states of minds of individual decision-makers.

The most important and entrenched variant of extending the basic framework as introduced so far is by so-called Bayesian conditionalisation, or updating. Indeed, Bayesian conditionalisation can also be viewed as part of the core decision theory, as marked by the fact that most of the above accounts are often called 'Bayesian' decision theories. In this method, new probabilistic information received by the decision-maker is integrated into the existing probabilistic beliefs by applying some variant of Bayes' theorem (Howson, 1997). Hence, theories of Bayesian updating are able to deal with cases where new information is acquired. This is arguably a vital extension to the static $\langle p, v\rangle$-representation, as it allows one to correctly model the states of minds of decision-makers who learn, who communicate with others, or receive cues from their environment. Recall the ex-

ample of the individual evaluating the prospects of going to the beach and staying home. Imagine she learns that her neighbours will be throwing a big party, or that she listens to the weather report. Conditionalising her beliefs on such new information will arguably be a better representation of her attitudes, if she is rational.

We will call the elements of decision-theoretic representations as introduced so far a 'standard' account. That is, a standard account combines an evaluation of prospects by beliefs and desires, given by the maximisation of a preference ordering over prospects represented by an expected utility function, with Bayesian conditionalisation to model the acquisition of new beliefs. We will now discuss whether such a standard account can be applied directly to analyse intertemporal decisions, and what kinds of extensions have been proposed in order to do so.

2.2.2 Intertemporal Decisions and Games

This section shows that a direct application of standard decision-theoretic approaches is not sufficient to provide full answers to the three problems of intertemporality in decisions and games raised in the introduction of this thesis. We discuss some extensions to decision theory that have been offered in the literature and suggest that while they answer some variants of the three questions posed here, further extensions to decision theory are needed.

Time Distance

The problem of time distance in intertemporal decisions arises when events associated with prospects are distributed over time. Indeed, there are intertemporal decisions, such as the decisions between the two dinners at different days, in which time distance plays a very important role.

To discuss how time distance can be analysed by applying standard decision-theoretic frameworks, firstly recall that the quantified representations of beliefs and desires are defined over prospects. A temporally extended prospect, such as a specific career path an individual can choose, is described by a set of all possible world histories that can be affected by the career path. Indeed, on this reading, this includes all possible consequences of this career path, such as what kind of life the individual will lead, what kind of people she will meet, and so on. Such a maximally rich description of prospects, however, is not without its problems. Once all consequences are spelled out, it leads to the conclusion that 'a person has

only one decision to make in his whole life. He must, namely, decide how to live, and this he might in principle do once and for all' (Savage, 1972, 83). To avoid such problems, decision theories usually make assumptions that permit them to characterise more isolated decision situations. Savage (1972), for instance, goes on to describe how decision theories based on such a 'lifetime-decision' assumption would not be applicable to practical decision-making, and indeed attempts to define 'small worlds' in which practical decisions can be analysed. This discussion of prospects illustrates that standard decision theories start from a rather general characterisation of decision-makers and prospects, without prescribing much structure.

How can time distance aspects of prospects be evaluated in decision-theoretic frameworks? In the standard accounts, intertemporal aspects are not modelled explicitly. That is, prospects can extend through time, and the decision-maker can adopt any attitudes to those, as long as those are within the confines of the structural assumptions that are needed to obtain expected utility representations. In particular, the decision-maker can take any attitude to time distance. That is to say, it is a matter of desirability attitude, or personal taste, how a decision-maker is influenced by distance in time. Therefore, decision theory does not assume that the present is inherently more, equally or less valuable than the future, or vice versa. If an agent is very patient, then the beliefs and desires that reflect this are just as admissible as the beliefs and desires of an agent who dislikes waiting.

In other words, decision theory does not offer structure for considering time distance separately. The beliefs and desires represent the agent's attitudes, and assuming specific attitudes to time that are separate from those is not part of standard decision-theoretic frameworks. However, there are many decisions in which the time aspect is of overwhelming importance, for instance those about handling climate change for collectives, or choosing a career path. A direct application of standard decision-theoretic frameworks as introduced above will not yield a detailed analysis of the intertemporal aspects of such decisions.

The above discussion suggests that in order to evaluate time distance, extensions to decision theory are required that can facilitate such an analysis. The multiple-self model proposed in the next section provide such an extension, offering a structure in which to model the deliberations of a decision-maker about temporal distance. We will discuss how it can be used to evaluate time distance

in Chapter 4. Other proposed extensions to decision theory, such as the concept of discounted utility in economics, will also be considered in detail in Chapter 4, and discussed in the context of the multiple-self model.

Interaction over Time

The problem of interaction arises when there is an interdependence between several decision-makers, such that the consequences of their decisions depend on each other. Rational interaction of decision-makers is analysed in game theory, which builds on standard decision-theoretic frameworks and furnishes additional structure and assumptions to consider the interdependence of players' choices. Interactions *over time* are analysed as dynamic games in the so-called extensive form, which will be introduced formally in Chapter 5.

The extensive form makes use of standard decision-theoretic accounts in order to model interaction over time; in particular, it is assumed that each player in the game is endowed with a utility function. Furthermore, the concept of belief updating is highly relevant, as dynamic games are marked by the fact that new information can become available at different stages in the game. More specifically, consider that interactions over time require extensions to standard frameworks of belief updading in two key respects.

Firstly, the beliefs involved in the characterisation of rational interactions are complex, as higher-order beliefs are needed to characterise a player's beliefs. That is, we are not only interested in the player's beliefs as such, but also in what she believes about her opponents, what she believes her opponents believe about her, what she believes her opponents believe what she believes about them, and so on. Here, the research programme of epistemic game theory offers us modelling devices that can characterise such higher-order beliefs (Brandenburger (2007), Perea (2011, forthcoming)). In Chapter 5, this approach will be will introduced formally.

Secondly, one of the most pressing questions in the analysis of interaction over time consists in the problem of accommodating seemingly contradicting beliefs, such as when players are faced with surprise information. For instance, a standard assumption in dynamic games consists in the belief in the rationality of opponents. However, the reasoning method of backward induction requires to entertain situations that could only arise due to an irrational move by an opponent (Stalnaker (1998). How to reconcile such surprise information with the

belief in the opponent's rationality has been a pressing problem in game theory. Binmore (1987)). In Chapter 5, we will give a characterisation of a new belief revision policy that deals with this problem.

More generally, considering belief hierarchies and surprise information are two problems of interaction over time that have led to extensions to standard frameworks. The multiple-self models that will be introduced in the next sections will help to further motivate such extensions, and will be used to interpret the sufficient conditions for backward induction developed in Chapter 5.

Temporal Dynamics

Temporal dynamics is perhaps the problem of intertemporal decisions that has received the most attention in terms of extensions to decision theory. Two families of extensions are particularly important: firstly, those extensions that analyse the problem of changing desires, by generalising standard decision-theoretic frameworks. Secondly, those extensions that deal with sequential decisions over time. We discuss each of those in turn and suggest that while they can treat a variety of problems associated to temporal dynamics, the description and explanation of dynamic inconsistency still requires further extensions.

Firstly, recall that Bayesian conditionalisation can model learning processes of decision-makers. Consider the example of choosing between the earlier or later dinner. For instance, in between the two possible dinners, a decision-maker could learn that the restaurant has received a damning review by an important critic, which could influence the attitudes of the diners. Temporal dynamics that are associated with the decision-maker receiving new information can hence be analysed with such tools. Applying such standard decision-theoretic tools in game theory and microeconomics more generally often comes with the assumption that the decision-maker's tastes are not supposed to change over time, and are not influenced by time in an explicit way. As Stigler and Becker (1977, 76) put it: '...one does not argue over tastes for the same reason one does not argue over the Rocky Mountains – both are there, will be there next year, too ...'. Embarking from this methodological proposition, Stigler and Becker (1977) then develop models of taste changes in which they are explained by belief changes. More generally, the application of decision-theoretic frameworks in economics is almost exclusively based on stable tastes.

However, temporal dynamics are not restricted to a decision-maker acquiring

new beliefs. It could also be the case that an agent changes her tastes or desires, rather than her beliefs. (We will henceforth use the terms taste and desire interchangeably). Recent generalisations of the concept of Bayesian updating also investigate an application of this method to desires, i.e. the value function (Bradley, 2007a, 2009a,b). In such applications of Bayesian updating, more complex and realistic cases of changes of decision-makers' minds can be accommodated for, such as changes of taste, for instance due to habituation, training and experience (Bradley, 2009b, 222f., for a list of examples). This account thus extends the standard decision-theoretic and Bayesian conditionalisation approaches and can deal with more complex temporal dynamics that are marked by taste change.

A related extension to standard frameworks in order to explain taste changes consists in introducing more structure in the model that yields the probability and value evaluation. Consider the recent model by Dietrich and List (2009), who introduce the idea that consequences can have different features, which can but need not become salient for the decision-maker. Hence, while there is only ever one $\langle p, v \rangle$-pair activated, a whole collection of those objects is possible, as and when different features of the world become salient, for instance when an individual grows up and develops different tastes to those he or she had as a child. In similar spirit, one can also make such an assumption directly about the decision-maker, such that it consists of a collection of those different $\langle p, v \rangle$-items in the background, as, for instance, the idea of avatars suggested by Bradley (2009b). Hence, if one assumes that each point in time is associated with more than one probability and value evaluation, considering collections of such evaluations – possibly indexed by points in time – makes it possible to extend on the analysis of temporal dynamics by standard representations.

A second family of approaches deals with a different aspect of temporal dynamics, namely that of sequential decisions. Sequential decision theory deals with the normative assessment of a particular kind of preference change, commonly referred to as dynamic inconsistency. This type of preference change refers to decision-makers which reveal contradictory preferences over time. For example, take the decision whether to go home after work, or whether to go to the pub. Imagine that a decision-maker knows that if he will go to the pub, he will drink too much and regret it the next day. Sequential decision theory deals with the normative assessment of such choices. On the one hand, proponents of so-called 'sophisticated' choice recommend to factor in the foreseen preference change in

the deliberations, recommending to go home. On this account, going to the pub (and regretting it afterwards) is brandished as naive or myopic (Hammond (1976), Steele (2007)). On the other hand, proponents of so-called 'resolute' choice recommend to find ways to prevent the momentary preference change from happening (such as McClennen (1990)). For a comprehensive overview and assessment of sequential decision theory, see Steele (2007).

Apart from the normative assessment, there is also the descriptive and explanatory question as to why real-world decision-makers often reveal dynamic inconsistency. That is, how can we describe and explain dynamically inconsistent preferences of decision-makers over time? Here, Schelling (1980, 1984) and Ainslie (1992, 2001) have proposed to draw on the metaphor of persons as 'multiple-selves' in order to analyse dynamic inconsistency, such that different selves have opposing preferences. The next sections presents a modelling approach to multiple-selves that is simpler yet achieves the same aims as the aforementioned ones, and makes transparent in what way the description of dynamically inconsistent decision-makers requires us to depart from normative decision-theoretic accounts, which will be discussed in detail in Chapter 6.

2.2.3 Extending Decision Theory

The above review has shown that standard decision-theoretic representations 'stay silent' on many aspects of intertemporality. This feature is by no means problematic in itself – indeed, part of the appeal of decision theory is the fact that its structure is simple and general. However, when analysing intertemporal decisions, extensions of decision theory that pay particular attention to the temporal dimension could widen its scope. If we were to take only a standard decision-theoretic approach to intertemporal decisions, then it would not be possible to discuss in greater detail in what sense discounting for temporal distance can be rational and how exactly it can be employed, how relaxing stability requirements in game-theoretic frameworks can be interpreted, and how to model dynamically inconsistent decision-makers. Yet, as the initial review of these problems in the introduction suggested, such topics are subject to considerable debate, and also of practical importance.

In order to accommodate the influence of time in decision theory, it is natural to nevertheless stay close to its framework. That is to say, the subjectivist character of decision theory need not be given up in order to model the influence of

time. In this context, it is interesting that Savage (1972) called his derivation of Bayesian decision theory 'personalistic' decision theory, and subjective Bayesianism as a whole was known as 'personalism' before terms like individualistic and subjectivist became more common (Teller, 1975; Zellner, 1982). This suggests that Bayesian decision theory is linked to the assumption of a decision-maker as a single individual. In fact, this feature of the theory is of great importance for its foundational role in economic theory and other social sciences that rest on methodological individualism. It is therefore natural to account for time in the context of normative decision theory in a subjectivist manner, developing an account of the internal temporal structure of the decision-making individual.

The multiple-self models introduced in this chapter are proposed to offer such a structure, which is intended as a minimal addition to decision-theoretic frameworks. Before introducing the multiple-self models, we review the rather diverse literature that has used the metaphor of the 'multiple-self' in various ways.

2.3 Multiple-Self Accounts

This section reviews accounts that employ the notion of multiple-self. Characterising persons as a collection of distinct and interconnected entities has been considered in many different philosophical traditions as well as in psychology, literature and economics. The main motivation of those accounts is to provide a more complex understanding of individuals. However, as ubiquitous as the presence of the multiple-self notion is, as elusive it has proven to be in terms of theoretical characterisation. This motivates the development of multiple-self models in the next section.

2.3.1 Elster's Review of Multiple-Self Theories

In a general sense, the idea of a multiple-self is that of a person consisting of several distinct yet interconnected entities. Elster (1986) provides a review of philosophical, psychological and economic theories of the multiple-self, whose accounts differ on a number of dimensions. We consider the comprehensive overview of such theories offered by Elster (1986), before discussing how the multiple-self model relates to the accounts in the literature.

The loosely integrated self. Elster (1986, 3) suggests that in many cases, multiple-selves 'turn out to be little more than failures of coordination and integ-

ration', such as an individual with beliefs and motivations that contradict each other, or beliefs and motivations that differ with regards to which realm of life they are applicable to. Indeed, once individuals realise such contradictions, they could resort to standard techniques such as revising their beliefs to resolve them. Hence, a 'loosely integrated self' might not have any more significance than a local or momentary departure from the idealised decision-maker as usually presumed in applications of Bayesian decision theories.

Self-deception and weakness of will. In the philosophical literature, multiple-selves are mostly associated with cases of self-deception and weakness of will. That is, the concept of several selves within one person is evoked in cases in which decision-makers act irrationally, act against their best interest, or hold conflicting attitudes (Elster, 1986, 6). There are two fundamentally different modes of inquiry regarding these problems: one concerns the possibility of its existence (Davidson, 1980, most prominently) and another one concerns its resolution (such as Ainslie (1992, 2001)).

Faustian selves. The famous 'two souls' within the *Faust*-character give the label for a type of multiple-self in which one part of the person has higher-order intentions whereas the other part has desires that clash with those. Such conflicts between the two Faustian selves can be superficial (i.e. potentially be resolved by deliberation) or very deeply seated, such as in prolonged inner conflicts. Schelling (1980, 1984) has proposed to view individuals in conflicts as two such opposing selves and has proposed to apply game-theoretic tools to model potential resolution of their conflict.

Hierarchical selves. Elster (1986, 11) also shows that it is possible to characterise the hierarchical nature of the Faustian selves more generally – and open the possibility for more than two selves in such a hierarchy within the person. On those accounts, the asymmetry between selves can be further characterised by differences in power, scope or relevance of the selves. For instance, considering the hierarchical approach of meta-preferences (or second-order preferences), a new type of preference is introduced that orders the preferences according to higher-order considerations. In Jeffrey (1983, 214), an agent can have a first-order preference that ranks smoking over not smoking. For his second-order preference, however, the same agent could prefer not to have that preference. Such models of second-order preferences are used to capture different types of motivations, i.e. those that concern betterness of alternatives and those that concern how an

agent would like to view the betterness of alternatives.

The Freudian legacy. Freud's theory provides the famous labels of the 'id', 'ego', and 'superego'. Elster (1986, 20) characterises these three concepts as 'agents' that are assigned different tasks in the person. He explains how these three agents map onto the tasks – or 'territories', in his terminology – of the 'conscious', 'preconscious', and 'unconscious'. Those concepts exemplify tripartite ontologies of persons and provide a metaphorical vocabulary for inner processes and dynamics, such as deeply rooted conflicts and their resolutions.

Split brain - split mind? The finding that the two hemispheres of the human brain can operate independently has, in theories of personal identity over time, given rise to many thought experiments that involve a split brain, which will be briefly discussed in the next chapter. Elster (1986, 23ff.) takes this as starting point to discuss whether 'cognitive compartmentalisation' of different degrees really does imply a divided self: he maintains that goals and motivations need not differ in a person who has abnormal communication between the two hemispheres of her brain (although this is possible). Cases in which cognitive compartmentalisation has such drastic effects seem to fall outside the scope of what a reasonably general theory of persons would want to cover.

Parallel selves. Elster (1986, 17f.) labels cases in which a person seems to enter a different state of mind with the notion of 'parallel selves'. That is, in cases of vivid imagination, such as daydreaming or in cases of concentrated cognitive effort, such as reading, the self of a person seems to be divided between two profoundly different worlds. Such consideration are important because they can help to characterise how person's imagine future situations and devise strategies accordingly, for instance against anticipated regret.

Homo oeconomicus and homo sociologicus. Elster (1986, 25f.) maintains that we can have different selves in private and in public, such that private utility-maximisation (according to an inner homo oeconomicus) can conflict with social norms and desired altruistic behaviour (according to an inner homo sociologicus). The divide between such motivations has been widely debated in a number of disciplines. For instance, in a series of papers, Akerlof and Kranton (2000) have explored the role of social identities for economics behaviour, attempting to bridge the gap between the traditional homo oeconomicus and homo sociologicus accounts.

Successive selves. Elster (1986, 13f.) also considers the case of successive

or temporal selves which is the focus of the multiple-self models introduced later, and discussed in greater detail in the next chapter.

The 'no-self' theory. Finally, Elster (1986, 28ff.) describes approaches that are extremely reductionist as those that deny that there is such a thing as a self, labelling those theories 'Neo-Buddhist'. In this category, he includes Hume (1739) and Parfit (1984) who endorsed 'disintegrating' views of individuals as vast collections of selves. In so doing, Elster concludes that they effectively deny that there is such a thing as the self.

In yet a further step to generalise the idea of a multiple-self, dropping even the assumption that there is a common person in the background of the multiple-self has been considered:

> 'A man is said to be the same person from childhood until he is advanced in years: yet though he is called the same he does not at any time possess the same properties; he is continually becoming a new person ... not only in his body but in his soul besides we find none of his manners or habits, his opinions, desires, pleasures, pains or fears, ever abiding the same in his particular self; some things grow in him, while others perish.' (Plato, *Symposium*)

Indeed, this first step to consider one person at different times really as a new person can be further developed into one where introspective processes of persons are compared to groups of distinct individuals. This idea already plays a prominent role in Plato's *Republic*, when he compares the inner structure of the Republic and that of the Soul (Pettit, 2003). Indeed, in the *Republic*, Plato regards the individual as state-like and the state as a super-individual. – While this view has important consequences for moral philosophy and political theory, it has remained metaphorical in the context of rational agency.

Now consider the following two dimensions of comparison of those accounts. Firstly, multiple-self accounts differ in how literally they take the notion of the multiple-self. Some theories use it as little more than a metaphor, whereas other theories go as far as associating multiple-selves with different physical entities and mental processes in individuals.

Secondly, the theories differ in the way they perceive of the different selves. Some theories explicitly assume a duality or tripartite of selves, others postulate a collective of selves and yet other theories conclude that from the notion of multiple-self follows that there is *no* self at all. Furthermore, some theor-

ies explicitly state distinct tasks for specific selves ('planning self', 'short-term interest self', etc.) while others allow do not introduce explicit distinctions. Additionally, theories also differ concerning the dimensions on which they perceive a multiplicity of selves: selves can be explicitly perceived of as successive, parallel, temporal, or as determined by contexts, social roles and so forth. Moreover, the relation or 'interaction' between selves is perceived differently, ranging from strict hierarchies, models of competition between selves, to complete equality of influence.

2.3.2 Towards Multiple-Self Models

The brief review of multiple-self accounts shows that the literature on multiple-selves is not only interdisciplinary but very diverse in its aims and scope. Furthermore, the terminology of the multiple-self seems to invite a metaphorical employment of the notion. While multiple-self accounts introduced in the previous section capture important intuitions about conflict and diverging motivations in a decision-maker, their disparate nature makes it hard to compare them to each other and to relate them to decision-theoretic representations of decision-makers.

The multiple-self models provided in the following section aim to improve on these deficiencies. In a general sense, the focus of the multiple-self models introduced below is to analyse the influence of time on decision-making by providing a simple structure which can be related to standard decision-theoretic approaches. Some of the above accounts, such as those that introduce successive selves and different social roles will be compatible with some of the models introduced.

Yet, not all of the above accounts will be compatible with such an approach. More precisely, the simple and general structure that we will introduce can be used in conjunction with standard decision-theoretic accounts. That is, we do not claim here to develop a new account of rational agency. Rather, we build on decision-theoretic account of rational agency and attempt to enrich it with multiple-self models.

2.4 Multiple-Self Models for Decision Theory

This section introduces multiple-self models as a tool to analyse the time dimensions of decisions. Firstly, we introduce the basic concepts of 'temporal selves' and 'connectedness', and present reductive and non-reductive interpretations of

these concepts. Finally, we discuss how multiple-self models relate to decision theory in greater detail.

2.4.1 Selves, Connectedness, and their Interpretation

The multiple-self model introduced here has three main components: temporal selves, connectedness, and their interpretation. Intuitively, temporal selves refer to the fact that decision-makers extend over time, and connectedness refers to the idea that many characteristics of decision-makers stay stable over time or only change very little. Finally, the interpretation of both temporal selves and their connectedness makes clear what we mean when we employ those terms.

The Simple Multiple-Self Model A simple multiple-self model is a tuple $M = \langle S, c \rangle$ where

- $S = \{S_0, S_1 \ldots, S_k\}$ is a set of temporal selves, drawn from some set Σ,
- $c : S \times S \rightarrow [0, 1]$ is a function that assigns degrees of connectedness to all pairs of selves $S_i, S_j \in S$.

Firstly, consider *temporal selves*. We can depict those selves as being drawn from some abstract set Σ such that each of the temporal selves in the set S corresponds to a point in time, as depicted in the following table.

Time	$\mathbf{t_0}$	$\mathbf{t_1}$	...	$\mathbf{t_k}$
Selves	S_0	S_1	...	S_k

Table 2.1: Temporal Selves in a Simple Multiple-Self Model

We can now ask in what sense the *temporal selves* are related to each other. The second component of the model is hence the idea that there is a degree of *connectedness* between the temporal selves. Connectedness characterises how strongly two temporal selves are connected to each other. For instance, for many decision-makers, it is a plausible assumption that each self is perfectly connected to itself, such that $c_{i,i} = 1$. Furthermore, for heavily idealised characterisations of decision-makers, the connectedness is perfect between all possible pairs of selves. Note that the assumptions of depicting temporal selves as a finite set and postulating a function in the above models are introduced for illustrating the kinds of

specifications of multiple-self models that will be employed in later parts – that is, the assumptions will be varying depending on the context of application.

Finally, consider that the temporal selves and their connectedness are the two concepts in this model that require an *interpretation.* The latter will require us to specify what kinds of objects temporal selves are taken to be and how we characterise their degree of connectedness. This, in turn, will motivate more specific discussions about how to obtain numerical values for degrees of connectedness. We now consider 'reductive' and 'non-reductive' interpretations. Reductive interpretations conceive of selves as sets of preferences and of connectedness as their degree of similarity. That is, they give interpretations that are closely related to concepts that are already contained in decision theory. Non-reductive interpretations conceive of selves as seats of a broad range of psychological features and of connectedness as capturing their degree of similarity. Before explaining those two kinds of interpretations in greater detail, we give an example to show how the multiple-self model relates to standard decision theories.

Consider the individual who chooses her career path. Take the evaluation of the prospect of studying for a postgraduate degree after having earned an undergraduate degree. This prospect is temporally extended, that is, there is a collection of possible events at different points, such as studying for a postgraduate degree, graduation, entering the labour market, and having a career based on having earned a postgraduate degree. Naturally, such a prospect can be subjected to a decision-theoretic analysis, by characterising the individual's attitudes to this prospect by her beliefs and desires. Indeed, the above model does not alter such a decision-theoretic analysis.

Rather, it adds a second step after considering such an analysis; namely, the multiple-self model allows one to express attitudes to time distance. That is to say, a decision-maker might not only form attitudes that can be characterised by beliefs and desires, but might also deliberate about the time distance that is inherent in the prospect. Precisely such deliberations can be modelled by temporal selves and connectedness. That is, a decision-maker could, due to introspection or by considering her past, form attitudes about her connectedness over time. It is then also possible that the decision-maker combines her beliefs and desires with connectedness to evaluate intertemporal prospects. That is to say, rather than changing decision-theoretic analysis, the multiple-self model offers an extension to it in order to better characterise attitudes to time distance.

Connectedness thus understood offers a variety of applications in the analysis of intertemporal decisions. Firstly, it can be used to motivate time discounting according to the general foundational framework developed in Chapter 4. Secondly, connectedness can give an explanation of a belief revision policy that underlies the new sufficient conditions of backward induction provided in Chapter 5. Thirdly, when using an extended multiple-self model, connectedness can be used in an analysis of dynamic inconsistency that is given in Chapter 6.

We will now turn to explain the different features of the multiple-self model in greater detail. In the next section, we discuss the non-reductive and reductive interpretations of temporal selves and connectedness, before discussing more complex variants of multiple-self models, and how they relate to decision theory.

2.4.2 Reductive and Non-Reductive Interpretations

This section explains the kinds of interpretations of temporal selves and connectedness in more detail. Firstly, consider reductive interpretations of temporal selves. In such interpretations, we can understand temporal selves as characterised by standard decision-theoretic representations; that is, as a set of preferences. Connectedness between temporal selves thus conceived can then be determined by the similarities of those preference sets.

Such preferences could range over prospects, as in standard applications of decision theories. Then, we would have to assume that every temporal self could perform an evaluation of full world histories. However, such a broad domain of preference does not have to be assumed in order to characterise the similarities and differences of temporal selves. It is practical to assume a smaller domain, such as specific events or consequences. Note that such consequences need not be prospects, but can also be simple propositions, such as 'I am eating icecream' or 'I am working hard', which different temporal selves may evaluate differently. In a general sense, all that is required in order to characterise temporal selves with decision-theoretic tools is a domain of objects which can be evaluated by all temporal selves.

A reductive interpretation allows us to give a more specific motivation of the connectedness function, as we can explain how the degree of connectedness can be quantified. This also suggests that the degree of connectedness is independent of the temporal prospect in question: it evaluates the degree to which preferences are stable over time, given a domain of preference on which such a stability

can be compared. That is, if temporal selves can be characterised by sets of preferences, and a common domain over which attitudes of temporal selves can be compared exists, then a connectedness function as introduced above can be given by comparing those different sets of preferences.

More formally, the function c measures the degree of connectedness between the temporal selves relative to some normalised measure of distance, difference or similarity between the attitudes of temporal selves. That is, $c_{i,j}$ is determined by a measure of difference between the attitudes of S_i and S_j. To give a simple example, the degree of connectedness can be obtained by a normalised Hamming distance between two sets.

Reductive Connectedness Measured by the Hamming Distance. Consider a small set of consequences $Q = \{a, b, c, x, y, z\}$ which is evaluated by S_i and S_j.

- S_i has the following preference ordering: $\{a \succ b \succ c \succ x \succ y \succ z\}$.
- S_j has the following preference ordering: $\{a \succ b \succ c \succ x \sim y \sim z\}$.
- The two orderings determine 15 binary relations and 3 of those are different, therefore the Hamming distance between those two sets is 3.[1]
- The connectedness $c_{i,j}$ can be determined by considering a normalised Hamming distance: $c_{i,j} = 1 - \frac{3}{15} = .8$

As briefly reviewed earlier, decision theories are well equipped to model changes in beliefs. Here, we assume that changes in preference between temporal selves as considered above are due to taste change. Reductive connectedness is hence a characterisation of the extent of taste changes between selves. According to such a procedure, we can establish similarity of preference between temporal selves, expressed as degree of connectedness c. We will explain in the next chapter of this thesis how such similarity of preference can be motivated by reductive accounts of psychological connectedness put forward by theories of personal identity over time.

Now consider non-reductive interpretations. Here, we can take temporal selves and connectedness as reflecting maximally rich descriptions of individuals that

[1]The binary relations between the pairs xy, yz, xz are changed and the ones between the pairs $ab, bc, cx, ac, bx, cy, ax, by, cz, ay, bz, az$ are unchanged.

exist over time. That is, we can take selves to be characterised by a wide range of physical and psychological features, and connectedness to reflect their degree of sameness over time. For the purposes of describing a decision-maker over time, we might nevertheless limit such a non-reductive description to a few salient features that are conceptually close to the kinds of characterisations in decision theory. That is, in a non-reductive sense, we can take temporal selves at each time – and their degree connectedness – to be determined by a broad range of psychological features, such as emotions, and feelings of empathy for other selves as well as memories. In order to describe such features more precisely, we will show in the next chapter of this thesis that conceptual content of theories of personal identity over time can be employed in this regard. More specifically, theories that characterise the connectedness between selves due to a continuity of private memories (Section 3.4.1), feelings of empathy (Section 3.4.2) or social relations (Section 3.4.2) can be used to describe non-reductive connectedness.

Note that adopting one of the reductive and the non-reductive interpretations does not change the way in which the multiple-self model relates to standard decision theory. In both interpretations, the model is applied in addition to a decision-theoretic analysis to capture time distance. In the reductive interpretation, time distance is associated with the degree of taste change in the decision-maker, and in the non-reductive interpretation, the changes that determine the degree of connectedness are more complex.

As such, the multiple-self model is rather coarse-grained, as it summarises attitudes to time distance in a single degree of connectedness that is independent of the prospect in question. In order to relax this assumption, we now consider more complex multiple-self models.

2.4.3 Dual Multiple-Self Models

This section considers another type of multiple-self model that will be relevant in some applications in the later parts of the thesis. Recall that in the above specification of the multiple-self model, a simple variant has been given, where there is one self at each point in time. Here, in order to consider more complex attitudes to time, we consider a dual multiple-self model in which there are two sets of selves at each point in time.

Indeed, a dual multiple-self model specifies two sets of temporal selves which can differ according to both their degree of connectedness and the interpretation

that is given.

The Dual Multiple-Self Model A dual multiple-self model of personal identity is a tuple $M_d = \langle A, C \rangle$ where

- A contains two personalities, labelled P and D that each consist of temporal selves, drawn from some set Π:

$$A = \left\{ \begin{array}{ccc} P_0 & P_{\ldots} & P_k \\ D_0 & D_{\ldots} & D_k \end{array} \right\}$$

- $C = \{c^P, c^D\}$ is a set of connectedness functions for the respective personalities.

According to such a dual model, it is also possible for a decision-maker to have a more complex structure to his identity. That is, we introduce two different temporal selves for each time point. Such a model allows us to consider deep ambiguities and conflict in the decision-maker's deliberations about intertemporal decisions. For instance, it may be possible that the decision-maker has different social roles that induce different attitudes to time, such that when an individual considers her role as mother of her children she has a higher degree of connectedness to her future selves than when she thinks of her role as a professional.

Such a dual model can also be seen as giving a 'disaggregated' version of a simple model with one row of temporal selves. However, for most of the applications in this thesis, variations of the simple multiple-self model already suffice to analyse intertemporal decisions. We will introduce a more detailed motivation of the additional structure in a dual-self model in Chapter 6.

2.4.4 Multiple-Self Models and Decision-Theoretic Representations

This section explains how the multiple-self models introduced above relate to decision-theoretic representations of individuals. First and foremost, note that the multiple-self models are proposed as *extensions* of decision theory. That is, they do not require a modification of existing decision-theoretic accounts. Rather, the models can be applied as a separate step in analysing intertemporal decisions, over and above a decision-theoretic analysis. Note that both steps do

not interfere with each other in relevant ways: it is still possible for the decision-maker to adopt any specific attitudes to intertemporal prospects as captured by beliefs and desires. In particular, the individual can, due to her desires or beliefs, still evaluate future consequences as equally, more, or less important as ones that are closer to the present. Yet, after having considered those attitudes, specific intertemporal attitudes can be characterised by connectedness.

Two Modes of Interpretation

To explain this in greater detail, consider two intuitively plausible uses of multiple-self models. Firstly, we can take such models to describe the deliberations of decision-makers about the intertemporal aspects of prospects. That is, a decision-maker can deliberate how well future selves are connected to her current self. In this mode of interpretation, the attitudes expressed by the degree of connectedness do not concern the evaluation of the intertemporal prospect, which are already given by beliefs and desires, but rather the kinds of changes that an individual considers relevant over time. The precise nature of those attitudes will depend on whether we assume a reductive or non-reductive interpretation of such deliberations.

Secondly, we can interpret such models from the theorist's point of view. On this reading of the three components of the model, they are a sparse representation of separate temporal attitudes of a decision-maker. This also makes it more plausible to introduce the kinds of similarity measurement procedures that we alluded to above, where preferences of different selves are compared.

These two modes of interpretation suggest that multiple-self models can extend the scope of decision theory if intertemporal decisions require us to analyse the temporal dimension separately. Intertemporal prospects such as a career path are marked by the fact that trade-offs made over time are a much more salient feature than the kinds of properties that can naturally be evaluated by beliefs and desires. Indeed, 'after' having evaluated decisions by beliefs and desires, the question might still stand what kind of attitudes to adopt to time distance or temporal dynamics. For those kinds of decisions, multiple-self models provide extensions to decision theories.

Correspondence

In the discussion of the multiple-self models, time indices were introduced to denote temporal selves (such as in a set of temporal selves $S = \{S_o, S_1, \ldots, S_k\}$). It has been tacitly assumed that such time indices relate the idea of different instances of decision-makers to time (such as in Table 2.1). In most modelling contexts, time indices can be commonly assumed and such a formalisation does not pose additional problems. Yet, if we are to interpret the multiple-self models as capturing what is relevant about time for decision-making, such time indices need further interpretation.

To make this step of interpretation explicit, we introduce the concept of *correspondence* between selves and a time-index. If such a correspondence holds, the temporal selves are associated with time points which means, in turn, that the evaluations they make are about the desirability and probability of objects that are also associated with the respective points in time. This correspondence to some externally given time-index is hence made explicit as an assumption. Variants of such a correspondence condition need to be assumed in order to attach a relevance to the time indices in the multiple-self models that goes beyond modelling. For instance, if desirability and probability evaluations are to be weighted according to which temporal self they concern, then such a correspondence condition will allow one to maintain that such weighting captures the influence of time on the evaluations. In the application of the multiple-self models in the next chapters, such a correspondence will be introduced explicitly as and when needed.

Separability

The multiple-self model requires us to accept a separability condition: just as decision theory implies separability of outcomes to draw a decision matrix and apply consistency requirements, an application of the multiple-self model implies a separability of times, most notably a separability of an agent at a time and the deliberation he engages in about his future selves.

Such a separability assumption is, in the context of analysing attitudes to time, not uncommon (for a detailed discussion, consider Broome (1991)). However, if we are to employ connectedness in conjunction with decision-theoretic concepts, the assumption of separability becomes stronger. That is to say, we have to endorse a temporal separability assumption concerning the beliefs and

desires of decision-makers, as connectedness is defined relative to such time points.

For instance, consider the possibility of applying degrees of connectedness as weights on specific consequences or events. In such applications, which will be discussed in detail in Chapter 4, goodness experiences by a low-connected temporal self is slightly devalued. However, in order to perform such weighting, the goodness itself also needs to be subjected to separability. – We will make such an assumption explicit as and when it is employed (and not already implicit in the frameworks that are amended with the multiple-self model).

2.5 Conclusions

This chapter has developed modelling tools to extend decision-theoretic representations of decision-makers. Embarking from a review of standard-decision-theoretic accounts, we have seen that those models stay silent on many crucial problems of intertemporality, such as how to evaluate time distance and how to model specific aspects of temporal dynamics.

Multiple-self models have been introduced to capture attitudes to intertemporality. Temporal selves and their connectedness can be used to give reductive and non-reductive characterisations of attitudes to time and intertemporality, and can be used alongside decision-theoretic evaluations of intertemporal prospects. Such multiple-self models can be seen as characterising an agent's introspection about her stability over time; that is, they give a degree of connectedness which signifies the similarities and differences between different temporal selves. We have also shown that such models clarify the use of the metaphor of the 'multiple-self', which has been widely used to suggest psychologically more realistic models of decision-makers. In a general sense, the multiple-self models present a structure that requires us to make explicit what kinds of phenomena we associate with intertemporal decisions, i.e. whether we analyse time distance, or more complicated problems associated to temporal dynamics.

Before discussing three problems of intertemporality in decisions and games, which benefit from applying the device of multiple-self models, we consider how key questions and distinctions in theories of personal identity over time cohere with the general structure of multiple-self models.

Chapter 3

Personal Identity over Time

Summary. This chapter shows how multiple-self models structurally cohere with key questions and distinctions in theories of personal identity over time. We suggest that those theories can be viewed as competing answers to three questions: (i) instances, the question of what is significant for a person to exist at a given point in time – this can be captured by the notion of temporal selves in a multiple-self model, (ii) persistence, the question of what is significant for a person to exist over time – this can be captured by the notion of connectedness in a multiple-self model, and (iii) criteria, the question of what establishes instances and persistence of persons – this can be captured by an interpretation of temporal selves and connectedness in a multiple-self model. That is, given a sufficiently rich specification of a multiple-self model, it can be motivated and constrained by specific theories of personal identity over time.

3.1 Introduction

Persons change over time: we grow up, adopt new attitudes, and vary our physical appearance. Yet, despite such changes in physical and psychological characteristics, we have the sense that we also stay the same persons over time: we retain ownership of past actions, and many of the aforementioned characteristics change only incrementally. Theories of personal identity over time elucidate our understanding of the seemingly contradictory nature of difference and sameness of persons. These theories aim to establish how and why a person at different times can still be the same person. This question has been considered in all philosophical traditions and many different accounts, theories and problems as-

sociated with it have been brought to the fore. This chapter reviews parts of this literature, limiting itself to accounts in contemporary analytic philosophy. There is no particular view or argument that will be argued for in the process of this review: rather, the goal is to identify those contributions in the literature on personal identity over time that can motivate and constrain multiple-self models.

This chapter aims to show that several key distinctions and questions in theories of personal identity structurally cohere with multiple-self models as developed in the previous chapter. For this, we discuss theories of personal identity in a tripartite framework. More specifically, we suggest that those theories can be seen as competing answers to the following three questions: (i) *instances*, the question of what is significant for a person to exist at a given point in time, (ii) *persistence*, the question what is significant for a person to exist over time, and (iii) *criteria*, the question of what establishes instances and persistence of persons. This framework will sometimes be referred to as a model of personal identity over time. We will show how this framework is compatible with other taxonomies and frameworks for theories of personal identity theories that have been proposed in the literature.

Recall that the previous chapter has formulated multiple-self models with three elements: temporal selves, connectedness, and their interpretations. The review of theories of personal identity will show that the notion of *selves* can capture the concern in personal identity theories to specify what *instances* of a person are, the notion of *connectedness* can capture the concern in personal identity theories to specify the *persistence* of persons, and the specific *interpretation* of connectedness and selves captures the concern in personal identity theories to specify a *criterion* of personal identity over time. In a colloquial sense, we might refer to this structural coherence as an isomorphism between the tripartite structure of multiple-self models and models of personal identity over time.

The structural coherence makes it possible to use the conceptual content of personal identity theories to motivate and constrain multiple-self models of personal identity over time. That is to say, we can use criteria of personal identity over time to give substantial content to the interpretation of selves and connectedness in multiple-self models. We will call models that are enriched in such a fashion 'multiple-self models of personal identity over time'. These models will be used to furnish additional insights into the three problems of time in decisions and games that are discussed in Part II of this thesis.

This chapter is structured as follows. Section 3.2 gives a broad historical overview of the main debates in theories of personal identity over time. Section 3.3 introduces the framework instances, persistence, and criteria in greater detail and shows how it is an appropriate framework for capturing theories of personal identity over time. Section 3.4 gives an overview of memory and psychological criteria of personal identity over time that can be used to motivate and constrain interpretations in multiple-self models of personal identity over time. Section 3.5 discusses problems of rational agency related to multiple-self models of personal identity over time. Section 3.6 concludes.

3.2 A Stylised History of Theories of Personal Identity over Time

The problem of personal identity over time has received widespread attention in all philosophical traditions and modes of inquiry. These contributions form a huge literature which is reviewed in Noonan (1989), Kolak and Martin (1991), Shoemaker (2008) and Olson (2008). Standard anthologies include Perry (1975a), Rorty (1976) and Martin and Barresi (2003). This section gives a brief historical account of how the most salient paradigms in theories of personal identity over time have emerged before analysing the accounts with regards to the three key concerns of personal identity theories alluded to in the introduction.

3.2.1 Plato and Descartes versus Locke and Hume

In the historical introduction to their anthology on personal identity, Martin and Barresi (2003) provide a broad division of the history of personal identity theories in analytic philosophy into three phases: firstly, from Plato up until Locke, secondly, from Locke to the 1960s, and thirdly, from the 1960s up to the present. Martin and Barresi (2003) admit and demonstrate that this three-phases view is rather coarse-grained, yet argue that it provides an understanding of how the main paradigms in theories of personal identity have emerged.

The first phase from Plato up until Locke is characterised by Martin and Barresi (2003, 6ff.) as being dominated by theories that present the so-called 'simple view' of understanding the self as being constituted by an immaterial and indivisible substance, such as the soul or the Ego. Plato first developed the notion of the soul as constitutive of the self in *Phaedo*, where he characterises it

as immaterial, indivisible and immortal. The concept is developed in discussions of human mortality, advancing the idea of the immortality of the soul. Thus, Plato's *Phaedo* does not aim to advance a complete picture of the psychological reality of the human mind and its persistence.[1] Indeed, in his review of different accounts of the soul, Lorenz (2009) maintains that

> 'Phaedo's conception of soul is narrower than our concept of mind... The range of activities (etc.) that the soul is directly responsible for, and which may be described as activities of the soul strictly speaking, is significantly narrower than the range of mental activities. It does not include all of a person's desires, nor need it include all emotional responses, or even all beliefs.'

This idea of a spiritual substance that makes up the essential identity of the self was taken up in accounts of Scholastic scholars and, notably, René Descartes (Martin and Barresi, 2003, 17ff.). On Descartes' account of personal identity, the concept of the 'Ego' fulfils the role of the soul in Plato. The Ego is disembodied which means that it is an immaterial substance:

> '...I thence concluded that I was a substance whose whole essence or nature consists only in thinking, and which, that it may exist, has need of no place, nor is dependent on any material thing; so that 'I', that is to say, the mind by which I am what I am, is wholly distinct from the body, and is even more easily known than the latter, and is such, that although the latter were not, it would still continue to be all that it is.' (Descartes, 1637, Part IV).

Taking the soul, the Ego or some other purely mental entity as constitutive of the self, it follows that personal identity is an 'unanalysable fact' (Noonan, 1989, 16). This 'simple view' of personal identity has been challenged by a family of theories that maintain that personal identity needs further characterisation, starting with John Locke.

The second historical phase in theories of personal identity over time, from Locke to the 1960s, is characterised by Martin and Barresi (2003, 24ff.) as mainly

[1]Note that Plato also developed a much more psychologically minded concept of the soul in the *Republic*, which depicts the soul consisting of reason, spirit and desire and provides the basis for a much more detailed account of the psychology of the human mind. Farther, in the *Symposium*, Plato developed what appears to be a relational view.

advancing the so-called 'relational view', i.e. understanding the self as changing process of physiological and psychological elements. In contrast to the 'simple view', which postulated a somewhat mythical idea of the self, such relational accounts lend themselves to further empirical analysis, such as collecting evidence for the sameness of psychological traits in a person. On Locke's (1694) account, the continuity in consciousness ensures that a person is identical at different times and not on the substance of either the soul or the body: "...wherein personal identity consists: not in the identity of substance, but, as I have said, in the identity of consciousness, ..." (Locke, 1694, Book II, ch. XXVII). (Note that due to changes in the use of those terms, Locke uses 'consciousness' here for what we now call 'memory').

Locke's view has marked the beginning of many proposals of criteria of personal identity over time that focus on different psychological traits that are said to be crucial in defining what makes a person at a later time identical to the one at an earlier time. Hume (1739) takes an even more radical stance in emphasising the relational perspective. Firstly, he is more forceful in denying the 'simple view', asking:

> 'what ... gives us so great a propension to ascribe an identity to these successive perceptions, and to suppose ourselves possest of an invariable and uninterrupted existence thro' the whole course of our lives?' ... [Persons are] nothing but a bundle or collection of different perceptions, which succeed each other with an inconceivable rapidity ... (Hume, 1739, Book I, Part IV, ch. vi)

Secondly, he also subscribes to the memory criterion of personal identity, maintaining that:

> '... memory alone acquaints us with the continuance and extent of this succession of perceptions, it is to be considered, upon that account chiefly, as the source of personal identity. Had we no memory, we never should have any notion of causation, nor consequently of that chain of causes and effects, which constitute our self or person.' (Hume, 1739, Book I, Part IV, ch. vi).

Discussing personal identity over time in terms of continuity of psychological features such as memory has yielded a number of distinct problems which are

discussed in recent accounts. Indeed, these early theories discussed so far already broadly foreshadow the lines of inquiry in more recent debates.

3.2.2 Contemporary Debates: Dualisms and Criteria

The third, contemporary, phase of personal identity theories can be roughly characterised by two families of developments. The first family of developments concerns a detailed analysis of distinctions along the 'simple versus relational view' dualism, advancing related, but more specific dualisms such as 'reductionism versus non-reductionism' and 'endurance versus perdurance'. The second family of developments concerns intricate debates about the correct criterion of personal identity, marked by the infamous methodology of thought experiments, such as the brain transplantation, fission and teletransportation cases.

We start with the first collection of developments that offer more detailed analysis along the 'simple versus relational view' dualism. The latter is closely related to the distinction between reductionist and non-reductionist accounts of personal identity over time, largely due to Parfit (1984). He maintains that theories of personal identity can be reductionist or non-reductionist: non-reductionist approaches share the concern to not limit the characterisation of personal identity to specific, empirically verifiable, factors like reductionist ones do. More specifically, reductionism endorses at least the first of the following two claims:

> '(1) that the fact of a person's identity over time just consists in the holding of certain more particular facts, (2) that these facts can be described without either presupposing the identity of this person, or explicitly claiming that the experiences in this person's life are had by this person, or even explicitly claiming that this person exists. These facts can be described in an *impersonal* way.' (Parfit, 1984, 210)

Non-reductionist accounts reject those claims and involve, in the terminology of Parfit (1984, 210), a 'deep further fact', for which the Cartesian Ego is an example. In contrast, reductionist approaches reduce personal identity to analysable entities, aiming at an empirically more precise yet conceptually more pragmatic grasp of what personal identity over time consists in and how it can be well described.

Many theories will, if they endorse the simple view, also endorse non-reductionism, such as in Plato's *Phaedo* and in Descartes. Conversely, the relational view is commonly aligned with reductionism, as in Parfit. This, however, is not always the

case. For instance, Locke endorses both the relational view and non-reductionism. As pointed out by Quinton (1975) and Uzgalis (2009), even though Locke does endorse a criterion-based, relational discussion of personal identity, he also holds a more complicated ontology of persons which distinguishes between *man*, *thinking substances* and *personal identity*. Due to this distinction, Locke does not reject the notion of the soul (as, for instance, Hume does), but merely argues against it as establishing personal identity. That is to say, on a narrow definition of personal identity, Locke characterises it reductively by memory but also holds a non-reductive concept of personhood. Similarly, Lewis (1983) also endorses a relational view and non-reductionism.

Consider next the dualism of 'endurance versus perdurance'. Many discussions of personal identity will make reference to persons existing at specific points in time, such as 'person A at time t_1'. Such time-slices of persons or person-stages can be taken as merely empirical time-indexed reference to an enduring person. However, persons can also be understood as collections of separate yet interrelated time-indexed entities. Already foreshadowed in Hume (1739), this idea has gained further traction in virtue of the thought experiments mentioned above, due to Parfit (1984), and more generally due to the fact that the intricate examples require a precise analysis which is facilitated by referring to different time slices or stages of persons. In this context, the question of the ontological status of those person-stages or time-slices of persons has been considered. In short, perdurance accounts view persons as consisting in four-dimensional 'space-time worms' – that is, three-dimensional, temporal parts (such as Parfit (1984), Lewis (1983)), whereas endurance accounts deny that such entities have any ontological significance (e.g. Shoemaker and Swinburne (1984)).

Naturally, theories of personal identity can be compared according to these three dualisms. In a simplified use of those dualisms, they do form two coherent collections; one advancing a 'unifying' and the other advancing a 'decomposed' picture of persons and personal identity over time. The unifying view would be held by theories that endorse all three of the simple view, non-reductionism and endurance. For example, Plato and Descartes can be characterised as endorsing all three of those views. A decomposed view would be held when endorsing all three of the relational view, reductionism and perdurance. This is indeed Parfit's position. However, as already mentioned, the dualisms do not need to coincide in such a fashion. For instance, Locke endorses the relational view, non-reductionism

and – partly because of his non-reductionism – does not strictly endorse either one of the concepts of endurance or perdurance. Shoemaker, on the other hand, agrees with Locke on the first two distinctions but also clearly rejects perdurance (Shoemaker, 1963; Shoemaker and Swinburne, 1984). Finally, Lewis (1983) endorses a relational view, non-reductionism and perdurance. Hence, introducing such distinctions has enabled one to be much more precise in analysing and comparing different theories of personal identity.

The second family of contemporary developments in theories of personal identity over time concerns different criteria of personal identity and thought experiments that demonstrate problems with such criteria. In the wake of Locke's initial development of the relational view, various candidates for criteria that fulfil the relational view have been discussed, such as continuity of the body, the brain, as well as relevant physical and different psychological features (Noonan (1989, 2–13), Olson (2008)). These will be reviewed and compared in detail later. In the context of contemporary debates regarding the plausibility of such criteria of personal identity over time, thought experiments have become an important tool (Martin and Barresi, 2003, 3), notably due to the ones put forward by Williams (1956), Williams (1970), Nagel (1971), Lewis (1983) and Parfit (1984). Such thought experiments, and their consequences, being by no means crucial to the task at hand in this chapter, will only be briefly revisited later.

This concludes a brief historical synopsis of important topics in theories of personal identities over time. The discussion already suggests that there are two key requirements of such theories: firstly, to align along a number of dualisms and secondly to endorse a specific interpretation that is upheld with regards to the positioning concerning those dualisms. This general picture will be further developed in a tripartite model in the next section. Some of the aforementioned concepts and problems will be further explained and analysed in this framework.

3.3 Three Problems of Personal Identity over Time

This section offers a systematic discussion of personal identity theories along the three dimensions of (i) instances, (ii) persistence and (iii) criteria. The aim of this section is twofold: firstly, we aim to demonstrate how this framework coheres with taxonomies and distinctions that have been proposed in the personal identity literature. This will be achieved by relating the framework to specific

accounts in the literature and re-describing the dualisms introduced in the previous section along the three dimensions. Secondly, we show how the threefold structure of the framework makes it possible to use conceptual content of theories of personal identity to enrich interpretations of selves and connectedness in multiple-self models.

Note that personal identity over time is a particularly challenging topic due to its complex nature, as it combines problems of personhood, identity and time – all of which raise many problems on their own. For instance, personhood raises the questions of how a person can be defined and how persons can be adequately distinguished from one another, from animals and objects. Concerning identity, a key debate concerns the infamous dualism between qualitative non-identity and numerical identity. Finally, few philosophical problems have been as contested as the nature and understanding of time itself. There are thus many different ways in which the problem of personal identity over time can be understood – depending on whether once focuses on one of the aforementioned aspects or whether one attempts to advance a specific all-encompassing view on the problem. These conceptual complexities motivate the introduction of the tripartite framework, in order to extract conceptual content from personal identity theories that is relevant for characterising and interpreting multiple-self models for the analysis of intertemporal decisions.

3.3.1 Instances, Persistence and Criteria

The framework of philosophical theories of personal identity over time proposed here focuses on three key concerns which are referred to as (i) instances, (ii) persistence and (iii) criteria of personhood. These three dimensions are intended to broadly capture the main concerns of theories of personal identity as introduced in the previous section. We first introduce the framework and then discuss later how the distinctions already introduced relate to it.

In the proposed framework, there are thus three important dimensions to theories of personal identity over time:

- Instances of a person at a time. Theories of personal identity over time will make some claim[2] about what they take to be significant about a person's

[2]Such claims could be implicit or even dismissive of the importance of the problem of instances (or one of the other concerns).

existence *at a time*. This will be referred to as the part of the theory that talks about *instances* of a person.

- Persistence of a person over time. Theories of personal identity over time will make some claim about what they take to establish for a person to exist *over time*. This will be referred to as the part of the theory that talks about *persistence* of a person.

- Criterion of personal identity. Theories of personal identity will make some claim about what they take to substantively or materially establish instances and persistence of a person. This will be referred to as the part of the theory that talks about a *criterion* for a person.

This minimal set of rather broad concerns about personal identity over time formulated in the framework stems from and is closely related to a number of similar proposals in the literature on personal identity over time that identify a set of questions or concerns, notably Perry (1975c), Parfit (1984), Olson (2008), as well as Quante (2007). For instance, in the introduction to his anthology, Perry (1975c) discusses personal identity along a dualism between qualitative non-identity and quantitative identity of persons (mapping onto instances and persistence) and its substantial interpretations. Many more fine-grained distinctions can be made, including an analysis of the level on which claims are made, e.g. whether the theories claim to make ontological, metaphysical or epistemic statements. These more fine-grained distinctions map on the basic three dimensions outlined above. For instance, introducing further distinctions with regards to demarcating the ontological relevance of claims, Parfit (1984, 202) raises four questions: (1) What is the nature of a person? (2) What is it that makes a person at two different times one and the same person? (3) What is necessarily involved in the continued existence of each person over time? (4) What is in fact involved in the continued existence of each person over time? In the above framework, Parfit's (1) maps onto (iii) criteria and Parfit's (2) maps onto (i) instances and (ii) persistence in the framework adopted here. Parfit's (3) and (4) enable him to discuss more specific concerns about the ontological status of persistence. More broadly, the overview of Olson (2008) mentions eight problems of personal identity, some of which directly map onto the three dimensions above, some of which provide more fine-grained distinctions. This suggests that the tripartite model proposed here is compatible with received frameworks in the literature.

The framework introduced above lends itself to a characterisation of the dualisms introduced in the previous section. We start with re-describing the 'simple versus relational view' dualism. In the new terminology, a theory of personal identity over time holds the simple view iff it postulates *exactly* one *criterion of personal identity over time, which* alone *establishes persistence, and instances are merely of descriptive or empirical significance.* This is the case for theories such as Plato's and Descartes' which endorse the soul or Ego as the sole criterion of personal identity which alone establishes persistence. Instances do not figure beyond an empirical referent in their theories, neither in the soul's characterisation, nor in the analysis of its significance. Concerning the relational view, in the new terminology, a theory that advances a relational view postulates *a number (commonly one) of criteria of personal identity over time, that can identify instances whose relations, in turn, can establish persistence.* This holds true for accounts such as Locke's, who identifies instances by the memory criterion and then asks whether there are suitable relations between instances (such as recollections of experiences) which establish persistence. The other dualism introduced in the earlier review will be re-described further below.

Note that it is neither argued that the framework captures all main concerns of theories of personal identity over time nor that it specifies necessary and sufficient conditions of personhood. Rather, it allows one to describe, identify and compare personal identity theories in a way that they can constrain multiple-self models as introduced in the previous chapter. Indeed the fairly general structure of the personal identity model of (i) instances, (ii) persistence, and (iii) criterion structurally coheres with multiple-self models that specify temporal selves, connectedness, and their interpretation.

The next two sections will further motivate and describe the three dimensions of the personal identity model, first discussing the relations between instances and persistence, and then focusing on the criteria.

3.3.2 Instances and Persistence

The relations between instances and persistence are at the heart of the dualism of 'endurance versus perdurance' that was briefly introduced earlier. Before discussing this dualism in detail, it is helpful to consider another, much more fundamental dualism that has been discussed in the literature: the so-called dualism of identity between qualitative non-identity and quantitative identity. Discussing

this dualism highlights how the concepts of instances and persistence adequately reflect key distinctions made in the literature.

In a first step, we remind ourselves that theories of personal identity over time can be seen as a specific way of talking about identity over time. That is to say, by analysing the problem of identity over time in general, fundamental constraints and categories can be found that also have to apply to the more specific problem of personal identity in some way. In theories that deal with the problem of identity, as reviewed in Sider (2000), Sider (2001), Lowe (2001) Noonan (2008), and Gallois (2008), the so-called dualism of identity and its possible resolutions play a major role. More specifically, theories of identity highlight the importance of addressing the dualism of identity as well as giving a substantive account such as a criterion or interpretation of the object whose identity over time is in question. For instance, Sider (2000, p.81) begins his discussion of recent work on identity by stating:

> 'Let us divide our subject matter in two. There is first the question of criteria of identity, the conditions governing when an object of a certain kind, a computer for instance, persists until some later time. There are secondly very general questions about the nature of persistence itself.'

To illustrate the dualism of identity, take the example of the identity of a physical object like a chair: a successful theory of identity over time will be able to make statements about how we are to understand the qualitative changes in a chair over time while it will also be able to make statements about how we are to understand the quantitative identity of the chair in question. We can call that the interpretation of the dualism of identity. Ideally, such an interpretation accommodates our twofold intuition that the 'same' chair can be of different quality, e.g. after furnishing a new upholstery or painting it, we still think it is the same chair in the sense of quantitative identity. Ideally, an interpretation of identity over time would also give adequate criteria of when we are no longer speaking about the same object but rather about two or more different objects (Gallois, 2008). Accordingly, theories of identity over time focus on answering two main questions, following the 'dualism of identity': one, how can there be qualitative non-identity over time and two, how can there be quantitative identity over time? Seeking an account of identity over time that allows one to answer both questions satisfactorily is the goal of any theory of identity, saying 'what

matters' when dealing with the dualism of identity.[3]

This discussion is intimately related to the the dualism 'endurance versus perdurance' discussed earlier. Insofar as instances and persistence are mere placeholders that signal the identity dualism, awaiting interpretation under a substantive criterion, they are unproblematic notions. However, even in the absence of a substantive criterion of personal identity over time, claims about the status of instances and persistence can be made that need to be supported. Consider the question of *how exactly* does persistence arise and in what way does it make reference to instances? Consider David Lewis' characterisation of endurance and perdurance, as cited in Lowe (2001, 127):

> 'something perdures iff it persists by having different temporal parts, or stages, at different times, though no one part of it is wholly present at more than one time; whereas it endures iff it persists by being wholly present at more than one time.'

Hence, taking persons to be *perdurers* relies on a four-dimensionalist ontology of persons where instances are interpreted as temporal parts that are existing over time, just as three-dimensional temporal parts. In contrast, taking persons to be *endurers* takes persistence as the fundamental ontological category and does not interpret instances of persons as temporal parts – at most, they are interpreted as empirical, observable instances of a person's life.

Hence, in the terminology of (i) instances, (ii) persistence and (iii) criteria, the dualism 'endurance versus perdurance' can be re-described as follows. Perdurance accounts maintain that *persons are collections of instances which can be persistent, and the persistence might be produced in virtue of some property that can be captured by a criterion.* This is a position that, for instance, Hume (1739) and Parfit (1984) subscribe to, as both regard persons as collections of some instances (which can be identified by memory or psychological criteria) and which, under enough continuity under the criterion of personal identity, are said to persist. In contrast to the perdurance view, endurance accounts maintain that *persons are persistent, possibly according to some criterion and unrelated to that*

[3]In the terminology of Sider (2000, 2001), we can say that the stages $S1$ and $S2$ belong to some continuing F iff ϕ, where $S1$ and $S2$ exhibit (i) qualitative non-identity, F exhibits (ii) quantitative identity and ϕ exhibits (iii) an interpretation. If we wish to focus on analysing qualitative non-identity, we provide an interpretation of $S1$ and $S2$ as being temporal parts of the continuing F; others regard $S1$ and $S2$ as different stages in the life history of the continuing F.

we have epistemic or even empirical access to instances of them. This position is, amongst others, endorsed by Plato and Descartes whose accounts start off from a persistence perspective, according to their notions of the soul.

The discussion of the dualism of identity suggests the relevance of discussing the interpretation of (i) instances and (ii) persistence of persons and the way in which temporal parts of persons are significant for personal identity over time. In this context, Lewis' point about the difference between discussing questions of 'identity' or 'similarity' is helpful (Lewis, 1983, 157ff.). Lewis questions whether discussions of personal identity are really about identity. In his view, problems of identity are straightforward – as no two things can ever be identical – while problems of similarity or sameness can be more intricate. He views the problem of personal identity as being one of similarity, and hence one of degrees of sameness. In such an understanding of personal identity, the metaphysical problems associated with endurance and perdurance become less pressing. It is indeed such a 'thin' understanding of instances and persistence as characterising degrees of personal identity, for instance in virtue of some criterion of similarity or continuity, that is sufficient to motivate and constrain multiple-self models.

3.3.3 Criteria

In order to discuss the third element of the framework, that of criteria, we now turn to discuss what kinds of substantive criteria of personhood have been proposed in the literature. As briefly mentioned in the historical overview, personal identity over time has been discussed with reference to variants and mixtures of physical and psychological criteria. The following criteria of personal identity are the most prominent ones and have either been taken to be the sole criterion of personal identity or combined and collated with other criteria (as presented and reviewed in Perry (1975a), Martin and Barresi (2003), Noonan (1989), Shoemaker (2008), and Olson (2008)):

- Thinking substance, the Soul or the Ego (Plato in *Phaedo*; Descartes (1637); Chisholm (1976)),
- Physical criteria, such as sameness of the body, the brain or somatic sameness (Williams, 1956; Nagel, 1971; Thompson, 1997; Snowdon, 1990; Olson, 1997, 2003),

- Sameness of memory, consciousness or *quasi*-memory (Locke, 1694; Hume, 1739; Shoemaker, 1959, 1963),
- Psychological continuity (Parfit (1984, 205ff); Noonan (1989)),
- Continuity of empathy, intentions, or narrative (MacIntyre, 1984, 1989; Taylor, 1989; Schechtman, 2001, 2005; DeGrazia, 2005).

As the above list suggests, many different criteria have been put forward by which one can say that a person stays the same person over time. For example, the bodily criterion of personal identity says that it is the same organism of a person that makes a person the same person over time. Other approaches argue that psychological connectedness is essential to being the same person over time.

Concerning those criteria, both reductionist and non-reductionist views have been proposed. In the context of the model of personal identity introduced here, the distinction between reductionism and non-reductionism can be re-described as a disagreement about the metaphysical scope of the above criteria. In the terminology of (i) instances, (ii) persistence and (iii) criteria, on a reductionist account of personal identity over time, *a number of empirical criteria (usually one) are employed to express facts about instances and/or persistence which, in turn, establishes personal identity.* Such a view is endorsed, for instance, by Parfit (1984). On his account of psychological reductionism, sameness of psychological traits is the criterion that can establish persistence which completely captures personal identity. In contrast, on a non-reductionist account, *personal identity cannot be fully established and captured by factual criteria of instances and/or persistence.* Such a view is endorsed by Descartes and Plato, as mentioned earlier. It also holds for Shoemaker (1963) and Shoemaker and Swinburne (1984): even though they endorse a relational view of personal identity, they do not think that the characterisation captures personal identity completely. Similarly, Locke maintains that there can be a persistent soul besides a characterisation of instances of memory.

Discussing different criteria of personal identity over time, Noonan (1989) makes an important distinction between their substantial and empirical interpretation. In his terminology, we can distinguish between 'constitutive' and 'evidential' criteria of personal identity. The debates briefly introduced in the previous chapter are significant and have attracted attention because they were taken to advance a substantive understanding of persons and their identity over time, in an

ontological sense. The substantive interpretation of theories of personal identity over time sees them as advancing necessary and sufficient conditions for viewing one person at a specific point in time as identical to another person at a different time, in the most fundamental sense. However, one can also view some of the theories and criteria as merely advancing an 'evidential' interpretation. This is especially plausible for theories that are relational, reductionist and/or endorse perdurance and makes those accounts easier to accept. Note that for the purposes of motivating and constraining interpretations of multiple-self models with theories of personal identity over time, such an evidential interpretation of criteria is already sufficient. That is, one can view theories of personal identity over time as providing simplified accounts of important features of personal identity over time without subscribing to the idea that they conclusively postulate metaphysical truths.

Before discussing in greater detail how criteria of personal identity over time can be used to motivate multiple-self models, we consider personal identity thought experiments in the framework of instances, persistence, and criteria.

3.3.4 Personal Identity Thought Experiments

This section suggests that the tripartite framework of the personal identity model does not imply that we are glossing over the intricate problems of personal identity encapsulated in thought experiments that have been suggested in the literature. Recall the fact that thought experiments and the problems that result from their discussion are an important methodological device in the literature on personal identity over time. In general, thought experiments are used to demonstrate how a specific criterion of personal identity can be shown not to hold in all cases, i.e. how it gives rise to a counter-intuitive conclusion or a paradox (Gendler (2000), Wilkes (1988)). That is to say, such thought experiments are used to expose the conceptual limits of different criteria of personal identity.

We consider a brief example of a thought experiment to demonstrate the descriptive accuracy of the tripartite framework. Examples of hypothetical 'fission' of persons, such as discussed in Nozick (1981) and Parfit (1984), can be understood as cases where a person divides into two (seemingly) numerically different persons and both of those are qualitatively identical to each other, as well as to the pre-person. Fission consists in manipulating important features of a person in a way that results in more than one candidate for being identical to the

former person: for example, we are asked to imagine that a person's brain is extracted from her body and split in half. The original body is destroyed and the two hemispheres of the brain are transplanted into two identical bodies. As a result of such manipulations, two persons are now plausible candidates for being identical with the former person according to a number of criteria: both persons can give an account of being related through, for instance, psychological features, and through sameness of the brain. Yet, we would not necessarily think of those persons as persistent with the pre-person from which the brain was extracted, and it would be hard to tell which one of the candidates has stronger relations with the pre-person.

We show how such thought experiments can be suitably re-described in the framework. Firstly, consider one personal identity model PI_{Brain} that gives personal identity as sameness of the brain. The above thought experiment shows that this criterion cannot account for some cases such as the above. Adopting a tripartite framework to characterise a personal identity theory that postulates sameness of the brain does not impinge on the validity of the above thought experiment.

Secondly, by invoking more than one personal identity model, we can even further re-describe the thought experiment to better understand its structure. Suppose there are, in the background, also candidate models of personal identity that advocate sameness of psychological features via PI_{Psych} and sameness of the body via PI_{Body}. Re-describing the example, it becomes clear that the problem of giving counterintuitive answers arises for PI_{Brain} (and PI_{Psych}) because sameness of the body breaks down, yet there is sameness according to the two other accounts. This re-description makes transparent that the thought experiment exposes a tension between different criteria of personal identity. More specifically, it is used to show that sameness of the brain or psychological features cannot account for some cases such as the above.

Note that the very nature of those thought experiments – and their methodological merit – stems from explicitly ruling out such rich descriptions of personal identity according to the three PI-models just introduced for additional explanation. Indeed, the thought experiments are usually used to show the limits of *one* specific criterion (in the above case, either one of sameness of the brain or psychological features) in giving an account of personal identity. All what we intended to demonstrate here is that the tripartite framework does not

Yet, as mentioned before, the main use of the framework is to make precise and explicit the conceptual content of theories of personal identity over time for using it in multiple-self models. For this goal, the descriptive accuracy of the framework is the main concern.

This completes the task of demonstrating that the model of personal identity is an appropriate framework to discuss theories of personal identity, as it is compatible with key distinctions, dualisms, and problems as proposed in the literature. We will now turn to a more detailed discussion of those criteria that will be used to motivate multiple-self models.

3.4 Criteria of Personal Identity over Time

Multiple-self models coheres with the structure of the framework of personal identity introduced here: temporal selves can be seen as giving a specification of instances, connectedness can give a characterisation of persistence, and their interpretation can give a criterion. This suggests that specific accounts of personal identity theories can be used to motivate and constrain multiple-self models.

This section considers what kind of conceptual content from personal identity theories can be used to enrich interpretations in multiple-self models. More specifically, this section reviews two categories of personal identity criteria that can both be broadly described as psychological criteria: those criteria that are close to the informational aspect of rational choice, such as memory and consciousness as well as those that are close to the valuational aspect of rational choice, such as preferences, tastes and empathy. For each of the two categories, reductionist and non-reductionist variants are discussed.

3.4.1 Memory Criteria

This section discusses the conceptual content of an important family of criteria of personal identity over time, namely those that appeal to some variant of a memory concept to establish personal identity over time.[4] Firstly, different memory concepts are reviewed, and secondly, both reductive and non-reductive uses of such

[4]Note that in the literature on personal identity over time, memory criteria are often referred to as constituting psychological criteria, or as belonging to the class of psychological criteria. Here, in order to establish possible conceptual interpretations of sameness and similarity to use in models of temporally extended decision-makers, we make a further distinction according to what kind of psychological features the specific criteria advance.

criteria are presented.

Memory, Consciousness, and Self-Knowledge

As briefly introduced earlier, on Locke's account of personal identity, the continuity in 'consciousness' ensures that a person is identical at different times and not on the substance of either the soul or the body: '...wherein personal identity consists: not in the identity of substance, but, as I have said, in the identity of consciousness, ...' (Locke, 1694, Book II, ch. XXVII). The mental features that Locke (1694) is interested in here are similar to what we now refer to as 'memory' – the recollection of one's own past actions and events. Noonan (1989) describes the different meanings that have been ascribed to the term 'consciousness' and maintains that Locke has used it in a strong sense: 'When one is 'conscious to oneself' knowledge of something is shared with oneself alone. In this use of the expression one may be thought of as a witness to one's own acts' (Noonan, 1989, 43). Locke endorses the consciousness/memory criterion not only to describe persistence but also to characterise its strength and scope, as he maintains that '...as far as this consciousness can be extended backwards to any past action of thought, so far reaches the identity of that *person*' (Locke, 1694, Book II, ch. XIXVII). Noonan (1989, 43) characterises the conceptual content of Locke's criterion as 'shared knowledge had by a present self of a past self's actions which Locke thinks of as constituting personal identity.'

Numerous authors after Locke, starting with Butler and Reid, and stated more concisely in Shoemaker (1959) and Perry (1975b), have pointed out that his account is subject to a circularity objection: the first-person account of memory as self-knowledge already presupposes personal identity. That is to say, if we apply Locke's memory criterion to establish personal identity, we ask: can a person, at a specific instance, remember to having been witness to her own acts at an earlier instance? Now, Shoemaker (1959) and Perry (1975b) argue that the candidates for such items of memory that could establish persistence need already belong to the person whose identity we want to establish, as we are asking for a person's *own* acts. That is, in order for the memory criterion to establish personal identity, we already have to identify the right kind of candidate memories by some other criterion (Noonan, 1989, 56ff.).

'Neo-Lockeans' have formulated the concept of *quasi*-memory (or *q*-memory, for short) in response to this problem. This is a more inclusive concept and

separates the first-person account of memory from the act of remembering, such that the latter does not presuppose personal identity. On this account, a person has q-memory of some experience such as an action if she remembers having had such an experience, *and* if her memory of the experience was caused in the right way by the experience she remembers (Shoemaker (1959), Noonan (1989, 144–162)). Thus separating the act of remembering from ascribing the memory to a particular instance makes it possible to use the memory criterion without circularity.

Hence, moving from an internalist account, such as Locke's, to an externalist one, such as q-memory, brings with it the qualifier that in addition to a criterion, an 'appropriate' causal link is needed in order to establish personal identity with the criterion. It is a separate discussion what constitutes appropriate causal links, and beyond the scope of this work. In a nutshell, the qualifier of 'appropriate' intends to rule out causal links that are invoked in thought experiments, such as brain transplants or teletransportation.

Memory: Reductionism versus Non-Reductionism

The memory and self-knowledge criteria introduced above can be interpreted in two fundamentally different ways. In a reductive interpretation of such criteria, it is maintained that they lend themselves to a propositional characterisation, whereas in a non-reductive interpretation, such a formulation is not taken to fully grasp what we mean by recollection of actions, thoughts and experiences. For example, take the past event of going to the theatre. In a reductive interpretation, the experience of this event can be summarised as a proposition, whereas a non-reductive interpretation would maintain that not all what is relevant about the experience of going to the theatre can be reduced to such a propositional description.

Locke's account is non-reductive as it is an internalist, or first-person account. However, it is possible to interpret Locke's memory criterion in a reductive sense, when dealing with specific, propositional items of recollection (such as: 'I went to the theatre and enjoyed it'). While this is still internalist, and therefore non-reductive, the kinds of memory items involved can be much more clearly defined. The Neo-Lockean concept of q-memory, on the other hand, is externalist, and therefore completely reductive. More recent contributions on memory criteria have focused on developing the Lockean and Neo-Lockean concepts just described;

on the one hand using more detailed characterisation of the concept of 'self-knowledge' (O'Brien, 2007; Evnine, 2008) to put forward reductionist memory criteria by drawing on the epistemic resources of persons, and on the other hand developing more detailed accounts of introspection and self-consciousness for non-reductive accounts (Cassam, 1999).

Summarising the above discussions on the different possible formulations of how *memory* can be relevant as a criterion of personal identity, we can identify a reductive and a non-reductive memory criterion. In a reductive memory criterion, memory is conceptualised as items of information, such as propositions that can be associated with a former instance of a person via an appropriate link. In a non-reductive memory criterion, memory consist in a private, and much broader sense of recollecting experiences which is irreducible to an informational account. In addition to a propositional, reductive characterisation of memory items, the introspection of a person constitute a fuller recollection of how it felt like to have the experience in question.

These two criteria can be used to motivate and constrain multiple-self models, for instance when endorsing an interpretation of connectedness between temporal selves as being due to memories. Accordingly, the two variants of memory criteria allow us to interpret the connectedness between temporal selves as either being due to shared memories between temporal selves. Whereas the reductive criterion maintains that memory connectedness represents empirical facts about the memories involved, the non-reductive criterion denies this as it endorses a both more private and broader notion of remembering.

3.4.2 Psychological Criteria

This section reviews the conceptual content of psychological criteria, in the narrow sense of continuity of preferences, tastes, emotions, empathy and narrative. As the distinction between reductionism and non-reductionism is more deeply entrenched in this part of the personal identity literature, we first review psychological reductionism and then discuss non-reductionist critiques.

Psychological Reductionism

The most important discussions of personal identity over time in terms of reductionist psychological connectedness can be found in Parfit (1984). Building on the four questions of personal identity mentioned earlier, he formulates his psy-

chological reductionism in two steps, first advancing a 'psychological criterion' and then introducing his concept of 'relation R'.

Parfit's Psychological Criterion. '(1) There is *psychological continuity* iff there are overlapping chains of strong connectedness.[5] X today is one and the same person as Y at some past time iff (2) X is psychologically continuous with Y, (3) this continuity has the right kind of cause, and (4) it has not taken a 'branching' form. (5) Personal identity over time just consists in the holding of facts like (2) to (4).' (Parfit, 1984, 207)

In setting up this criterion, Parfit is permissive with regards to the conceptual content of psychological features – he specifically includes memory in the list of psychological features (Parfit, 1984, 220f.), yet is also adamant that it covers continuity of beliefs, desires and character traits. The latter, more inclusive interpretation of psychological features is explicitly endorsed by him for his more general criterion 'Relation R'.

Parfit's Relation *R*. Psychological connectedness and/or continuity with the right kind of cause (any cause) (Parfit, 1984, 215).

This more general criterion fulfils two roles: one is to further push psychological reductionism by allowing 'any cause', such as those that are used in thought experiments, to establish psychological connectedness and continuity. A second role is to generalise the content of what is understood as psychological features, explicitly including propositional attitudes, such as preferences, tastes and beliefs. That is to say, on this account, we can understand instances of a person establishing persistence if there is a similarity of psychological features, and if this similarity comes about by a cause that we are willing to accept as establishing persistence. There has been a large debate about the possible nature of such causes (Dancy, 1997), with much of it parallel to the debates about memory and q-memory. Hence, since Parfit's Relation R also provides an externalist account (via the permissiveness of the nature of causes that establish persistence), his criterion is also limited to – and in fact proposed as — constituting a reductive account.

[5]Parfit distinguishes between connectedness and continuity. The first term means similarity between instances of a person and the second term means that there are a overlapping layers of connectedness.

Reductionism and Non-Reductionism

Psychological reductionism as endorsed by Parfit (1984) has been criticised for a number of reasons (Dancy, 1997). Here, we deal with conceptual critiques that maintain that psychological reductionism fails to capture what we should take as relevant about the psychological features of persons. Almost all of the critiques can be taken as advancing the point that in order for reductive psychological continuity to hold, it must be enabled or produced in some way. That is to say, the features of persons captured by psychological reductionism are the results of much more complex processes that are unduly neglected in psychological reductionism.

There are two particularly relevant families of critiques in the context of modelling decision-makers as temporally extended persons: firstly, there are critiques that focus on attacking psychological reductionism on 'classic' grounds, maintaining that in order for psychological reductionism to hold, some fundamental psychological mechanism or capacity must be in place, be it a soul, an Ego, a continuity of empathy, or some sort of continuing mental life. Secondly, there are critiques which argue that psychological features of persons are produced by and closely related to external relations, such as those to other people or to the physical world, and those which see psychological features as ingredients in a much richer and complex narrative of personal identity.

Concerning the first critique, it can be summarised as an appeal to the 'commonsensical intuition of essential self-unity' (Belzer, 2005). As alluded to above, this intuition can be spelt out in two different ways in this context. One is to go back to the accounts of the simple view, according to which the soul or the Ego provides underlying self-unity. Another one is to give a fuller account of the mental life of a person, that does not amount directly to a rejection of the relational view, but insists on the presence of further mechanisms and capacities such as empathy or sympathy between instances of a person. For example, on the account that has been developed by Schechtman (2001, 2005), even though a person has changed (or will change) drastically with regards to her propositional attitudes, she could still have the feeling of psychological continuity, out of an understanding of how that instance of her person at a different time has enjoyed (or will enjoy) completely different things.

Concerning the second critique, Quante (2007) maintains that external relations, and specifically social ones, are vital in understanding persons, personhood and personal identity over time. In this view, the social nature of persons is a

deeply entrenched feature of the human condition, and as such it can be said to contribute to what we understand ourselves to be, shaping our psychological features. MacIntyre (1984), MacIntyre (1989), and Taylor (1989) also maintain that psychological features are produced by and embedded in the 'narratives' of person's lives. Reductionist accounts of psychological features at best focus on the results of such complex narratives. In particular, they fail to recognise the underlying mechanisms by which persons change and persist over time. Note, however, that reductionism is not incompatible with endorsing the second critique – it just gives a much more sparse characterisation of personal identity than the second critique endorses.

There are hence a variety of conceptual intricacies to the continuity of psychological features, in particular their very subjective and complex nature, that have led many authors to believe that in order for such features to establish persistence, a non-reductive account is needed. Perhaps less so than in the contrasts between reductive and non-reductive memory, the two variants of psychological criteria do not seem to be in strong opposition to each other – it is certainly plausible to endorse a reductionist account on grounds of methodology without dismissing that there is an underlying non-reductive mechanism that produces those features.

To summarise the above discussions, psychological features provide persistence criteria that can be used to motivate and constrain multiple-self models. Here we identify a reductive and non-reductive psychological criterion. In a reductive psychological criterion, psychological features are seen as specific traits such as tastes or preferences that can be associated with former instances of a person via an appropriate causal link. In a non-reductive psychological criterion criterion, further mental features are required in addition to reductive ones, such as empathy, external relations or the narrative of the life of the person in question.

These two criteria can be used to motivate and constrain multiple-self models, for instance when endorsing an interpretation of connectedness between temporal selves as being due to their psychological features. Accordingly, the two variants of the psychological criteria allow us to interpret the connectedness between temporal selves as either being due to similarity in their psychological features.

3.5 Multiple-Self Models of Personal Identity over Time

This section briefly summarises the multiple-self models of personal identity over time developed in both the previous chapter and the present one. In particular, we suggest a common terminology for the interpretations of temporal selves that will be used throughout the remainder of this thesis.

In Chapter 2, we have introduced the notions of temporal selves and connectedness, and given an example of their formal structure. We considered a set of temporal selves and a connectedness function that gives degrees of connectedness between pairs of temporal selves. Furthermore, we pointed out that those objects need an interpretation. We considered a reductive interpretation which conceives of the degree of connectedness as measuring similarity of tastes between pairs of temporal selves, and a non-reductive interpretation which endorses a broader range of features. In the remainder of the thesis, we will keep the terminology of temporal selves and connectedness, yet adopt the more general terminology of *psychological connectedness* for the reductive interpretation, and *empathy connectedness* for the non-reductive one.

In this chapter, we suggested that theories of personal identity can be used to motivate and constrain multiple-self models. That is to say, the interpretations in multiple-self models can be drawn from theories of personal identity over time. Indeed, *psychological connectedness* can be seen as giving a particular version of a reductive psychological criterion of personal identity. That is to say, we will from now on speak of psychological connectedness between temporal selves for a reductive interpretation. Similarly, *empathy connectedness* coheres with non-reductive psychological criteria, as discussed in an earlier section of this chapter.

Note that the above discussion would also permit to formulate a reductive and non-reductive memory interpretation of connectedness. However, for simplicity, we will focus on using the interpretations of psychological and empathy connectedness in Part II of the thesis. More generally, the above criteria can provide substantial interpretations of what exactly is captured by the formal structure in multiple-self models. We will henceforth refer to multiple-self models whose interpretations are given by one of the above criteria as 'multiple-self models of personal identity over time'. That is, such models will follow the structure of multiple-self models as outlined in the previous chapters, yet their substantive interpretation is compatible with conceptual content of theories of personal identity over time.

In general, the discussion so far yields a threefold motivation of multiple-self models of personal identity over time: firstly, we can motivate their application as an enrichment of decision theory in order to model the temporal dimension of decisions, as argued in the previous chapter. Secondly, this chapter has shown that the structure of multiple-self models also makes accessible conceptual content from theories of personal identity over time, which can motivate the idea of connectedness between temporal selves. Thirdly, we can motivate their application by the kinds of insight they will allow us into time and rational decision-making – this will be demonstrated in the next three chapters.

3.6 Conclusions

The concern of this chapter was to identify a structure that both connects reasonably well with existing theories of personal identity over time in two respects: one, capturing the most important traits of these theories to flesh out their differences and similarities and two, providing a structure that allows one to model such traits in order to motivate and constrain multiple-self models. Having looked at some of the most important proposals in the literature of personal identity over time, the 'personal identity triple' of instances, persistence and criteria aims to express key questions and distinctions in the contributions in order to make them available for multiple-self models.

Concerning both the distinction between evidential and substantial interpretation as well as regarding the distinction between similarity and identity, the *metaphysically weaker* understanding can be adopted in the following. This is due to the fact that in order to inform multiple-self models, an evidential understanding of the degrees of similarity of persons over time is already sufficient. This does not preclude assigning a greater significance, i.e. it is indeed possible to understand the models as providing substantive metaphysical foundations for the identity of changing decision-makers; yet, this is not necessary in order for the content of personal identity theories to have conceptual significance.

We have shown that many key dualisms and distinctions in theories of personal identity over time can be adequately described in a tripartite framework that structurally coheres with multiple-self models. This, in turn, makes it possible to consider the conceptual content of theories of personal identity over time in such models. We refer to those models as 'multiple-self models of personal identity

over time' as they are capable of capturing substantive criteria that have been offered in the literature. Those enriched multiple-self models – in addition to their capabilities to extend decision theories to analyse the temporal dimension of prospects – are hence grounded in accounts of how persons both change and persist over time.

As with the modelling device of the multiple-self models, the enriched models are not required as premises for the following discussions of three particularly interesting problems of time in decisions and games. Yet, as we will attempt to show in the second part of this thesis, for each of those problems, multiple-self models of personal identity over time offer us additional insight into the role of intertemporality in decisions and games.

Part II

Three Problems of Time in Decisions and Games

Chapter 4

Time Discounting

Summary. This chapter investigates how time discounting functions analyse temporal distance in intertemporal decisions. We identify two goals that theories of time discounting may have: one, postulating a correct time discounting function, and two, offering an accurate underlying conceptual motivation. We proceed by presenting a general representation framework for time discounting which outlines the requirements that well-founded time discounting functions have to fulfil. This general framework is used to analyse both existing accounts of time discounting, as well as Parfit's dictum of time discounting because of a weak connectedness to future selves. More generally, the requirements for time discounting theories developed here demonstrate that time discounting factors are restricted in the kinds of conceptions they can express.

4.1 Introduction

It is standard practice in the analysis of intertemporal decisions to introduce weightings that reflect the value given to the temporal dimension of a prospect. Such weightings are performed by time discounting factors that make goods in the far future less valuable than those in the near future. Famously, time discounting is a heavily contested concept (Loewenstein and Elster, 1992). There are two key problems: Firstly, there is no consensus on the correct functional form of discount factors, in particular, the properties of the discount rate that is often used in such functions are contested (Frederick *et al.*, 2002). In its most basic use, the discounting rate remains the same regardless of how far prospects extend through time. This most commonly used form of discounting originated

with Samuelson (1937) and is called exponential discounting, due to the shape of the value function it induces. More recently, 'hyperbolic' discounting, initially proposed by Ainslie (1975), postulates a declining discount rate, based on empirical evidence. Secondly, it is contested what kind of conceptual interpretation of time discounting is the right one: can it be explained by time impatience, attitudes towards risk and uncertainty, delay perception, or preference change? Each of these two key problems of time discounting can be understood both descriptively and normatively. Regarding the latter, it is often questioned whether time discounting is justified at all. Indeed, philosophers tend to deny the justifiability of time discounting (e.g. Sidgwick (1907), Rawls (1971), Broome (1991), Broome (1999)). Despite this, in the spirit of Ramsey (1928), time discounting is deeply entrenched in standard economic modelling:

> 'It is assumed that we do not discount later enjoyments in comparison with earlier ones, a practice which is ethically indefensible [...] we shall, however, ... include such a rate of discount in some of our investigations.'

The aforementioned normative and descriptive debates concerning time discounting have generated much disagreement, as reviewed by, for instance, Loewenstein and Read (2003). This fact renders scientific and policy debates about intertemporal decisions, such as those related to pension systems, public investment and climate change, deeply challenging. In order to clarify the concept of time discounting, this chapter asks the following question: How can we make sense of time discounting factors; what do they measure and represent? In other words, how can we meaningfully assign numbers to time points that can be used as weights for goodness evaluations of consequences that are associated with those time points?

In order to address this question, this chapter investigates the construction principles of time discounting functions. Note how this concern is different from the question that asks how to correctly evaluate intertemporal prospects in a general sense. Intertemporal prospects can raise a number of complex questions. Consider again the example from the introduction of this thesis of an intertemporal decision about whether to go out for dinner tomorrow or rather next week. Evaluating the intertemporal prospect of the dinner next week raises many issues, such as whether the dinner next week is an executable plan, whether fellow diners can be trusted to turn up, whether there will be regret for not having gone

earlier, and so on. Rather than considering the full array of those kinds of issues, the concern of this chapter is to clarify the exact role time discounting can play in the evaluation of such intertemporal prospects. We will show that, given certain assumptions, time discounting can contribute time-indexed weights to evaluating the time distance aspect of intertemporal prospects. Returning to the example, time discounting factors can be used, once goodness evaluations about the dinner are formed, to weight the expected goodness that this prospect provides. Yet, the explanatory scope of those weights is severely constrained by the assumptions that are required for their construction. In a nutshell, time discounting functions can serve as a coarse-grained evaluation of the influence of temporal distance on the evaluation of intertemporal prospects. Understanding the precise confines of the concept of time discounting is a key desideratum of the following analysis. The remainder of the introduction gives an overview of the structure of this chapter.

In Section 4.2, existing theories of time discounting are critically reviewed. Time discounting theories postulate time discounting functions that ascribe weights to time points, such that the present is assigned the unit weight and future time points are assigned weights in the real interval $(0, 1)$, with time points in the far future given smaller weights than those in the nearer future. We discuss the proposals of exponential and hyperbolic discounting, which introduce further restrictions on time discounting functions. Furthermore, different proposals for the conceptual motivation of time discounting are reviewed. We suggest that time discounting theories can be seen as competing answers to two questions: (i) what is the correct time discounting function, and (ii) what is the correct conceptual interpretation of time discounting factors given by such functions? These two questions can be asked on both a normative and a descriptive level, which yields four problems of time discounting. Such a division into four problems along the lines of functional form, conceptual interpretation, normative, and descriptive occurs also in other areas of enquiry, such as expected utility theory. In such contexts, representation theorems play a crucial role in clarifying answers to the four problems, as they provide a framework in which properties of functions and their conceptual interpretation can be specified.

Section 4.3 critically reviews existing representation theorems for time discounting. We start by giving an introduction to measurement-theoretic frameworks and highlight the crucial role of representation theorems. Frameworks and

theorems of representation are available for exponential discounting (e.g. Samuelson (1937), Koopmans (1960), Fishburn and Rubinstein (1982)) as well as for hyperbolic discounting (e.g. Strotz (1956), Manzini and Mariotti (2007), Halevy (2008)). Crucially, these representation theorems are obtained by assuming specific interpretations of time discounting. For instance, psychological notions like time impatience are invoked, objects like time preferences are integrated into existing theories, or other phenomena such as delay perception, attitudes towards risk and uncertainty, or preference change are used to motivate time discounting. We suggest that these frameworks treat aspects of intertemporality and goodness in a deeply entangled way, making it difficult to compare their relative merits in sufficient detail. Thus, in addition to the four problems of time discounting raised in Section 4.2, there is a more fundamental problem with time discounting theories, which lies in the absence of a representational framework for time discounting that would allow us to separate the evaluation of goodness and intertemporality to gain a precise conceptual understanding of time discounting.

Section 4.4 develops general foundations of time discounting that initially separates the the evaluation of intertemporality and the evaluation of goodness. For this, we firstly give a general definition of a time discounting function as the target of the representation. From the perspective of the representational theory of measurement, each weight that such a function gives has to be a numerical assignment to some salient qualitative property that is associated with a time point. We give a representation theorem that shows how a general time discounting function can be constructed that fulfils this requirement. Crucially, this result is obtained without a pre-commitment to any conceptual view about time discounting: indeed, this representation framework states general measurement-theoretic conditions for the construction of well-founded discounting functions. These conditions make transparent the fact that – from a measurement-theoretic perspective – any time discounting theory needs to endorse a numerical representation of some qualitative evaluation of properties that can be associated with time points. We also show how specific time discounting functions can be recovered by introducing further constraints within the general framework.

Section 4.5 discusses that the four problems of time discounting in the general framework. That is, the problem of the functional form of time discounting and the conceptual motivation of time discounting are discusses both descriptively and normatively. It is shown that the general framework renders explicit the regularity

conditions that are required in order to construct time discounting functions that are both formally and conceptually well-founded. We also reconsider the time preference theories of discounting, which are frequently used in economics to motivate the concept of 'discounted utility', in the general framework.

Section 4.6 discusses two specific proposals for time discounting. Firstly, the general framework is interpreted with the dictum of Parfit (1984) that 'my concern for my future may correspond to the degree of connectedness between me now and myself in the future.' In other words, the general representation theorem is interpreted as capturing the idea of connectedness in the multiple-self. Different interpretations of connectedness between selves are compared with regards to their plausibility of motivating time discounting. Indeed, by formalising Parfit's claim, we can revisit objections to it posed by Williams (1970) and Elster (1986). In a second step, we show that Parfit's claim supports a particular interesting interpretation of time discounting by the rate of preference change between temporal selves in a person. We show that if such a rate is constant, exponential discounting can be derived from the degree of preference change. This novel derivation of exponential discounting from preference change highlights the usefulness of the general framework of representation.

Section 4.7 concludes that the foundations of time discounting developed here make explicit what kind of assumptions are required in order to construct time discounting functions. Those assumptions delineate the evaluation of time distance from the kinds of evaluations that are captured in theories of utility and probability. This makes precise the confines of the role of time discounting for the evaluation of the time distance aspect of intertemporal prospects. Indeed, the framework is instrumental in distinguishing time discounting from other concepts of that raise more complex problems of intertemporality, such as interaction over time, temporal dynamics, or plans.

4.2 Time Discounting

This section reviews standard accounts of time discounting, such as exponential and hyperbolic discounting theories, and characterises the main debates and foundational problems associated with them, posing four problems of time discounting. This sets the scene for discussing the representation theorems of discounting theories in the next section and the general representational framework

developed and applied to those theories in the later sections of this chapter.

4.2.1 Time Discounting Functions

Discounting, in its most general meaning, is the lowering of the value of an object, good or prospect for a specific and separate reason. For instance, shops may lower the price of goods if a consumer purchases a large number or a specific bundle of them, policy-makers may disregard the opinion of someone with a vested interest, and individuals are often less affected by the suffering of people that they do not know personally, and so on. *Time discounting* refers to the practice of weighting the value of an object, good or prospect with a factor that is related to the time of their occurrence. For instance, the prospect of getting a piece of fruit in a month's time will be assessed by first evaluating the goodness of receiving the fruit and then applying a discounting factor that reflects the fact that the fruit will only be received in a month's time. Thus, a time discounting factor is a weight that is supposed to capture the influence of the time dimension of prospects on their evaluation. More generally, time discounting factors are time-indexed weights which are applied to evaluations of goodness.

To understand how time discounting factors are commonly used, consider time discounting in the context of a standard $\langle p, v \rangle$-framework. Take the prospect of having dinner today and assume that an agent evaluates this prospect in a way that reflects her rational preferences, for instance $V(\text{Dinner}) = 10$. Commonly, this can be taken to reflect an all things-considered subjective evaluation of the worthiness of today's dinner for that agent. Now consider a variation of the example in which the dinner will take place tomorrow. In this case, we can take the information that the dinner takes place tomorrow as forming a completely new prospect and consider a new all things-considered subjective evaluation of the agent. However, another possibility is to take the initial evaluation of the dinner today $V(\text{Dinner}) = 10$ and introduce a factor that reflects the fact that the dinner is held tomorrow, assuming that the dinners are otherwise identical. This is the idea of time discounting: the initial evaluation of a prospect is multiplied by a *time discounting factor* $D(t)$. Such a time discounting factor $D(t)$ is usually assumed to take a value between 0 and 1, thereby diminishing the initial value ascribed to prospects. To continue with the example, the *discounted* value of having dinner tomorrow is calculated by multiplying the initial evaluation $V(\text{Dinner}) = 10$ with the time discounting factor for tomorrow $D(\text{tomorrow})$. If the latter is,

say, $D(\text{tomorrow}) = .98$, then the discounted value of the dinner tomorrow is $DV(\text{Dinner tomorrow}) = 9.8$.

Discounting factors can be given by discounting functions. Those functions assign numerical values to points in time. In a general sense, a discounting function can be described as a mapping from a set of time points $T \subseteq \mathbb{R}$ to the real numbers, i.e. as a function $D : T \rightarrow \mathbb{R}$. T can also be discrete, and in most applications it is assumed to be a set of non-negative integers, with 0 denoting the present and all other points representing points at future times. Either a finite horizon (i.e. $T = \{0, 1, 2, \ldots, t_{max}\}$) or an infinite horizon ($T = \{0, 1, 2, \ldots\}$) can be adopted. The number in $\mathbb{R}$ that is assigned to a time point by a discounting function is then used as a discounting factor for value that occurs at that point in time. For instance, analogous to the example above, if x_3 is a consequence indexed by the point in time it occurs at ($t = 3$), then its discounted value $DV(x_3)$ is obtained by weighting its initial evaluation $V(x_0)$ with the discounting factor $D(3) \in \mathbb{R}$ such that $DV(x_3) = D(3)V(x_0)$.

As a matter of convention, time discounting usually results in weighting future value slightly less than the same amount of these objects in the present or without discounting. For instance, in the aforementioned example, discounting usually leads to $V(x_0) > DV(x_3)$. Furthermore, discounting factors are usually lower for points in time that are further away. This reflects the idea that goodness at later times should be discounted higher than goodness at earlier times. Hence, most discounting functions are decreasing, such that $D(t) > D(t + 1)$. The range of the discounting function is usually restricted to a real interval such as $(0, 1]$. Negative discounting factors would result in negative values and discounting factors larger than 1 would increase the value assigned to future consequences. While such values for discounting factors are not logically impossible, most discounting functions do not include such values. Again, this is a matter of convention, and reflects the idea that weighting goodness with a factor that is determined by its time of occurrence results in a slight devaluation of the goodness.

A general discounting function D can hence be understood as a decreasing mapping from a set of time points T to the real interval $(0, 1]$ such that $D(0) = 1$. This is not intended to rule out the possibility of endorsing a more general discounting function such as $D : T \rightarrow \mathbb{R}$. Rather, such a function adequately reflects the common ground of many discounting function proposals and some of the conventions discussed above. Indeed, most time discounting functions in the

literature as reviewed in the following section are a special case of this function, offering more specific restrictions on what values discounting factors can take. Note that those specific functions have two different roles: one is to give a more specific rule for assigning numerical values to time points than the above function and the second role is to facilitate an interpretation of those numerical values.

These two different roles lead indeed to the two key questions about time discounting: firstly, what is the correct functional shape of a time discounting function? Secondly, what is the conceptual motivation for time discounting? The next two sections will review how these questions have been answered in existing discounting theories.

4.2.2 Exponential and Hyperbolic Discounting Theories

The most important discounting functions are exponential and hyperbolic ones. We firstly discuss exponential discounting, followed by hyperbolic discounting.

Exponential discounting functions introduce a constant discounting factor δ which is used to calculate the discounting factor for each point in time. That is, an exponential discounting function D_e can be given by a mapping from time points to a real interval such that $D_e(t) = \delta^t$, $0 < \delta < 1$. In most derivations of exponential discounting, the constant δ is given by a constant 'discount rate' $r \in [0,1]$ which relates to the discounting factor as follows: $\delta = \left(\frac{1}{1+r}\right)$. Hence, frequently, exponential discounting is described directly by $D_e(t) = \left(\frac{1}{1+r}\right)^t$. This is indeed the case for most standard applications of exponential discounting in economics. We will refer to this particular variant of exponential discounting as 'constant-rate exponential discounting', as it is also possible to obtain a constant discounting factor δ by employing other concepts than a constant rate. Conceptually, the discount rate reflects the time preferences of a time impatient agent. In economics, constant-rate exponential discounting is most commonly known as the discounted utility (DU)-model.

To avoid confusion about the different concepts involved in time discounting, we will clarify the meaning of the different technical terms introduced so far. The *discounting factor* is the number, assigned to a point in time by a discounting function, which is used for weighting goodness evaluations of intertemporal prospects. Such a discounting factor can be determined in a number of ways. One frequently employed method of obtaining discounting factors is by introducing the concept of a *discount rate*. Discount rates $r \in [0,1]$ are also called per-period (or

time-point) discount rates, as they are taken to reflect the weight that is attached to $t+1$ in t. That is to say, the concept of the discounting factor is conceptually more general than the discount rate, as the former can be determined in a number of ways (this point is also prominent in the influential review of time discounting theories by Frederick *et al.* (2002)). Discount rates are one specific, yet widely employed way to determine discounting factors. Indeed, many conceptual and normative debates discuss questions about time discounting in terms of 'choosing the correct *discount rate*' rather than discounting factors. This is due to the fact that many crucial differences between concepts of time discounting can already expressed in this slightly easier, yet less general, terminology: for instance, just as one can ask what a discounting factor represents conceptually, one can ask the same question of discount rates. Even more importantly, by introducing formal conditions on the behaviour of the discount rate, crucial differences between discounting theories can be expressed. For example, in exponential discounting, r is *constant* to reflect that such weights are equal between any two time-points. Hence, exponential discounting is often referred to as 'constant-rate discounting'. In contrast, many theories of 'hyperbolic discounting' can be (partly) described as endorsing a *declining* discount rate. The latter theories will be introduced below, including those that combine the discount rates with parameters that capture delays, deviation from constant-rate discounting, and error terms to obtain discounting factors. We will continue to use the more precise language of a discount factor that is given by a discounting function. Apart from greater clarity, there are two reasons for this: firstly, the ultimate goal of any time discounting theory is to give the correct discounting factor for each point in time under consideration. The discount rate is only one possible ingredient in this exercise. Secondly, there are theories that do indeed combine other parameters with discount rates to obtain discounting factors – it would hence be unduly narrow to only discuss discount rates.

The constant-rate exponential discounting model introduced above is due to Samuelson (1937). Interestingly, Samuelson (1937), Samuelson (1939) and Koopmans (1960) did not endorse constant-rate exponential discounting. Rather, they intended the derivation of exponential discounting to be a mathematically interesting result without great empirical or normative significance. Despite this, exponential discounting has become the standard method of time discounting in economic theory. The most important formal property of exponential discounting

is that it preserves the utility function, even when engaging in time discounting. Indeed, this very property lies at the heart of the enduring normative appeal of exponential discounting, as it rests on a representation of time preference as discounted utility (the representation will be discussed in the next section). Descriptively, the mathematical tractability of exponential discounting and the formal parallels to the marginal rate of substitution have made exponential discounting attractive (Frederick *et al.*, 2002). Yet, empirical evidence has amassed that questions the descriptive accuracy of constant-rate exponential discounting.

In this context, hyperbolic discounting has emerged from empirical study of how real-world agents discount for temporal distance (Angeletos *et al.* (2001), Frederick *et al.* (2002)). In such empirical studies, it has been found that real-world agents are 'myopic', in the sense that time differences in short horizons are perceived as more relevant than time differences in longer horizons. So-called 'hyperbolic' discounting functions capture this phenomenon as they decrease more drastically than exponential ones for short horizons, i.e. near the present. This results in more time discounting for short horizons than in exponential discounting. For longer horizons, some hyperbolic discounting functions behave similarly to exponential discounting functions while many decrease less drastically than exponential discounting which results in less time discounting for larger horizons.

Many variants of functions that capture the idea of myopia have been proposed (for an overview of the actual empirical studies that have led to the different discounting functions, see the reviews by Frederick *et al.* (2002), and Loewenstein and Read (2003)). Since hyperbolic discounting is informed by empirical research, there are a number of proposals that each captures a variety of data. The following functions determine a discounting factor by delays, discount rates, constants and/or factors such that discounting factors are generally smaller than in exponential discounting for earlier times.

- Discounting for delay: $D(t) = \frac{1}{t}$, where t equals the length of delay (Ainslie (1975), Ainslie (1992), Ainslie (2001)). This function results in no discounting for the next period and a steep decline of the discounting factor for the following periods. Note that the delay function assigns the unit weight to the present as well as to $t = 1$, and is only strictly decreasing thereafter. On the conceptual level, there are no time preferences introduced; all that matters for time discounting is the perception of the delay.

- Discounting for delay and discount rate: $D(t) = \frac{1}{(1+rt)}$, where $r > 0$ is the

discount rate and t the delay (Herrnstein (1981) and Mazur (1987)). This function behaves similarly to exponential discounting for the near future and discounts the far future less. In this theory, the consideration of time preferences capturing time impatience is combined with the consideration of how delays are perceived.

- Generalised (hyperbolic) discounting: $D(t) = \frac{1}{(1+\alpha t)^{\gamma/\alpha}}$, where $\alpha > 0$ measures how much the function departs from constant rate discounting and $\gamma > 0$ is a parameter related to time preferences (Loewenstein and Prelec (1992), Laibson (1997)). With the behaviour of this function depending heavily on the specific values of the two parameters, most hyperbolic discounting functions can be written as a special case of this function. Most of the empirically relevant specifications of the parameters result in discounting the near future more than exponential discounting and in discounting the far future less than exponential discounting.

- Quasi-hyperbolic discounting:

$$D(t) = \begin{cases} 1 & \text{if } t = 0, \\ \beta\delta^t & \text{if } t > 0. \end{cases}$$

where $0 < \beta < 1$ can be constant or decline as t increases and δ^t is the exponential discounting function (Phelps and Pollak (1986), Laibson (1986), Laibson (1997), Barro (1999)). This discounting function captures the idea of hyperbolic discounting in a much simpler way than many other functions as the weighting factor β can capture how much the discounting deviates from exponential discounting.

Figure 4.1 displays the graphs of the aforementioned discounting functions.[1]

[1]The graphs in Figure 4.1 are based on the following functions:

- Exponential discounting: $D(t) = \delta^t$, where $\delta = .8$ (i.e. given by $\frac{1}{1+r}$ where $r = .25$),
- Hyperbolic discounting for delay: $D(t) = \frac{1}{t}$,
- Hyperbolic discounting for delay and discount rate: $D(t) = \frac{1}{(1+rt)}$, where $r = .25$
- Generalised (hyperbolic) discounting: $D(t) = \frac{1}{(1+\alpha t)^{\gamma/\alpha}}$, where $\alpha = .7$ and $\gamma = .9$
- Quasi-hyperbolic discounting: $D(t) = \begin{cases} 1 & \text{if } t = 0, \\ \beta\delta^t & \text{if } t > 0. \end{cases}$ where $\beta = .8$ and constant, and δ as for exponential discounting.

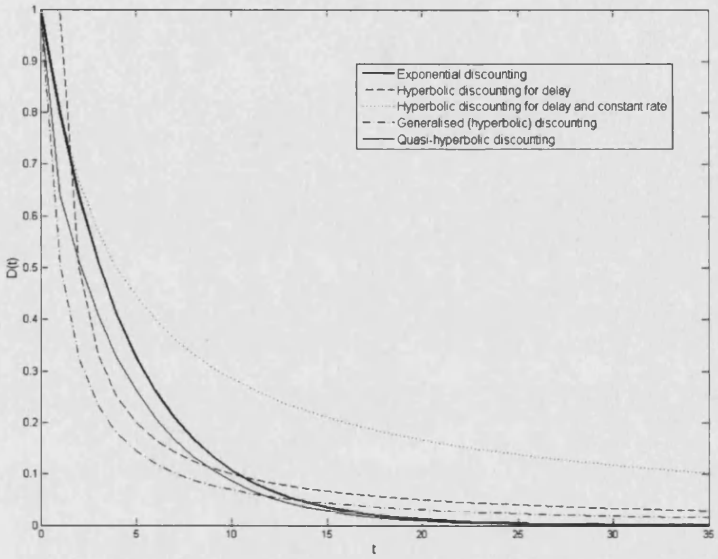


Figure 4.1: Discounting Functions

Prominently, exponential discounting weighs each subsequent time period with the same rate of discount r which results in a constant discounting factor δ of which the t-th power is taken to obtain the discounting factor for every point in time. This results in an exponential shape of the function. Most of the hyperbolic discounting functions yield smaller discounting factors for short horizons than exponential discounting. The notable exception is delay discounting, for period 1: note how the hyperbolic discounting for delay adopts a time horizon, such that for the delay of one time period, there is no discounting at all (i.e. $D(1) = 1$) and a steep decline of the function for the following times (indeed, $D(5) = .2$). Further note that for a longer horizon, some of the hyperbolic discounting functions yield less discounting (i.e. larger discounting factors) than exponential discounting.

The debate about the correct shape of time discounting functions concentrates, by and large, on these two proposals, with exponential discounting on the one hand and the family of hyperbolic discounting theories on the other hand.

4.2.3 Conceptual Motivations for Time Discounting

The above review of time discounting theories has concentrated on the specific shape of the discounting function that those theories endorse. Here, we briefly look at what kind of conceptual motivations have been discussed to underpin time discounting. The literature on possible conceptual motivations for time discounting is much more diverse than the literature on time discounting functions, and as a consequence, a whole host of motivations has been proposed to value future consequences less than present ones.

As mentioned earlier, time discounting is employed to weight goodness evaluations of prospects that extend through time. When discussing conceptual motivations for time discounting, the question arises how discounting factors are used. In other words, before answering the question 'what is the conceptual motivation for time discounting?', we first need to answer the question 'the time discounting of *what*?'. More specifically, to what kind of evaluations can time discounting factors be applied? Are those weights used to discount future monetary value, future natural resources, future utility, or future happiness? This question, as pointed out by Broome (1991) has led to a lot of confusion in discussions of discounting. For simplicity, he contrasts the discounting of monetary value on the one hand and utility on the other hand. Indeed, Broome (1991, 44) goes so far as to say that there is 'more misunderstanding than disagreement' between philosophers and economists, asserting that typically, economists do not employ time discounting for well-being and utility whereas philosophers focus on the justifiability of the latter. The widespread use of the standard DU-model in economics shows that Broome's ascription of economists as being mostly concerned with discounting future monetary value is not quite correct. However, the distinction between discounting future monetary value versus future utility is a useful one. In the remainder, we will concentrate on the discounting of utility and consequentialist goodness evaluations more generally.

There are a number of competing conceptual motivations for discounting future utility (or goodness evaluations), such as:

- Time impatience (or time preferences) (discussed by, for instance, Samuelson (1937), Koopmans (1960), Lancaster (1963), Fishburn and Rubinstein (1982)),

- Delay perception (discussed by, for instance, Ainslie (1975), Ainslie (1992),

Ainslie (2001), Laibson (1997), Ok and Masatlioglu (2007)),

- Risk and fundamental uncertainty about the future (discussed by, for instance, Weitzman (2001), Gollier (2002), Halevy (2008)),
- Preference change (discussed by, for instance, Strotz (1956), Parfit (1984), Laibson (1997), Frederick *et al.* (2002)), and
- Interaction between selves in a decision-maker (discussed by, for instance, Thaler and Shefrin (1981), Ainslie (1992), Ainslie (2001), Fudenberg and Levine (2006), and Xue (2008)).

Starting off again with constant-rate exponential discounting, the canonical interpretation of the constant discount rate r used in these theories rests indeed on the idea of *time preference.* More specifically, r is interpreted as the constant *rate of time preference.* The concept of time preference, in turn, is supposed to capture the idea that agents' *time impatience* plays a major role in the subjective evaluation of intertemporal prospects. Indeed, the concept of time impatience was at the very heart of the beginnings of the time preference theories of discounting: Frederick *et al.* (2002) point out that for many of the precursors of time preference theories, like Böhm-Bawerk, Fisher, Jevons and Pigou, the concept of time impatience was widely taken to be psychologically plausible and central in developing their theories of capital and interest. However, these authors have offered different and complex interpretations of how time impatience arises, forming an 'amalgamation of various intertemporal motives', according to Frederick *et al.* (2002, 355). For instance, Frederick *et al.* (2002, 353) cite John Rae's *Sociological Theory of Capital* (1834) in which he maintains that:

> '[t]he actual presence of the immediate object of desire in the mind by exciting the attention, seems to rouse all the faculties, as it were to fix their view on it, and leads them to a very lively conception of the enjoyments which it offers to their instant possession.'

Eugen von Böhm-Bawerk gives a characterisation of time impatience as consisting of underestimating future wants in his book *Capital and Interest* (1889), cited in Frederick *et al.* (2002, 354):

> 'It may be that we possess inadequate power to imagine and to abstract, or that we are not willing to put forth the necessary effort, but

> in any event we limn[2] a more or less incomplete picture of our future wants and especially of the remotely distant ones.'

This appear to echo the dictum of 'weakness of imagination' which Ramsey (1928) credited as producing time discounting. In the same vein, in *The Economics of Welfare* (1920), Arthur Pigou characterised time preference as arising 'from a type of cognitive illusion' (Frederick *et al.*, 2002, 354):

> '[...] our telescopic faculty is defective, and we, therefore, see future pleasures, as it were, on a diminished scale.'

While these authors endorse slightly different explanations of how exactly time impatience arises, they all offer these explanations to suggest that time preference reflects time impatience which they take to be a deeply rooted psychological fact. On the basis of those early conceptual considerations regarding the role of time impatience, Samuelson (1937) was the first to formally derive exponential discounting from time preferences. Note that, as highlighted by Frederick *et al.* (2002, 353), assuming that time preferences capture time impatience is a considerable conceptual simplification, when compared to the more complex discussion of time impatience by Böhm-Bawerk, Fisher, Jevons and Pigou.

As mentioned when introducing the hyperbolic discounting functions, there are a variety of concepts endorsed in hyperbolic discounting theories, including time preference, delay perception, risk and uncertainty, as well as preference change. Since hyperbolic time discounting theories of have been formulated as a result of empirical study, they are aimed at capturing the data of those studies, and indeed aim at predictive accuracy. It his hence not surprising that a variety of constants and factors determine those more complex functions. Indeed, the constants in the generalised and quasi-hyperbolic discounting functions are difficult to underpin conceptually. Still, from an explanatory point of view it is also plausible that real-world agents' attitudes towards the future depend on a variety of factors. However, the mixture of conceptual motivations behind hyperbolic discounting functions does not provide as straightforward motivations as with time preference theories. An exception from such conceptual complexities is the idea of delay perception in hyperbolic discounting for delay (given by $D(t) = \frac{1}{t}$) as only the idea that agents perceive of time differences in the near future more drastically than in the far future is invoked here. In addition, Ainslie (1992),

[2] *to limn.* to depict, or to picture.

Ainslie (2001), Ainslie (2005) supplies a theory of bargaining between temporal selves to underpin the theory of delay perception behind discounting for delay.

Hyperbolic discounting has also been motivated by the idea that both risk and uncertainty, as well as preference change are associated with distance in time. For instance, Weitzman (2001), Gollier (2002) and Halevy (2008) consider how time-indexed probability functions and risk evaluations can influence motivations for time discounting. Note, though, that probabilities have also been used to motivate exponential discounting (such as a constant probability that a decision-maker's life may end, (Mas-Colell *et al.*, 1995). In the context of hyperbolic discounting, Halevy (2008) considers how time impatience can vary, and establishes a dependence between time impatience and the perception of risk: present bias that is typical for hyperbolic discounting weakens when the immediate becomes risky. Preference change theories of time discounting motivate time differences with changes in the propositional attitudes of agents. In those theories, the future goodness evaluations of agents are discounted with their diminished present credibility due to changes in preferences, as suggested by, for instance, Strotz (1956) and (Frederick *et al.*, 2002, 389). Less formally, Parfit (1984) also suggests time discounting because of changes in preferences. The latter proposals have also been dubbed 'multiple-self' accounts of time discounting (for instance in the Frederick *et al.* (2002) review), suggesting that the present self evaluates prospects from her perspective and discounts the evaluations of future consequences to reflect that her future selves might have changed preferences. Furthermore, in Thaler and Shefrin (1981), Ainslie (1992), Ainslie (2001), Fudenberg and Levine (2006), Read (2006), and (Xue, 2008), decision-makers are explicitly assumed to be multiple-selves to discuss intertemporal decisions and time discounting. As mentioned in the initial review of the multiple-self literature in Chapter 2 of this thesis, those authors assume that there can be more than one self at a time (in most of those contributions, a decision-maker is assumed to consist of a far-sighted 'planner' self and short-sighted 'doer' selves) and attempt to discuss intertemporality in a wide sense, commenting on time discounting, problems of dynamic consistency, planning, and the formation of second-order beliefs by decision-makers about those problems. These proposals will be discussed in more detail in Chapter 6 of this thesis.

With the exception of only some theories (such as the complex motivations for time impatience by the precursors of the discounted utility model, and some

of the multiple-self accounts), the motivations underpinning time discounting do not form conceptually rich theories. In particular, it is hard to see how the conceptual motivations provide arguments for restricting discounting functions to exponential and hyperbolic ones. Moreover, in order to compensate for the absence of such accounts, many theories – especially hyperbolic ones – appeal for a combination of the aforementioned motivations. While the variety of conceptual motivations in discounting theories may not be problematic in itself – indeed, it can be argued that such an entanglement is necessarily involved in an adequate description of the phenomenon – it makes it difficult to compare those theories to one another.

4.2.4 Four Problems of Time Discounting

The question which of these aforementioned theories of time discounting is the correct one, has not been resolved (overviews of the debate can be found in Frederick *et al.* (2002) and Loewenstein and Read (2003)). In general, the theories introduced in the previous sections can be compared with regards to two questions: firstly, does a given theory of time discounting provide the correct time discounting function? This is the question of providing the correct discounting factor. Secondly, does a given theory of time discounting provide the correct interpretation of time discounting? This is the question of providing the right kind of conceptual content that underlies the time discounting.

These two questions can be discussed in two fundamentally different modes; namely, descriptively or normatively. In a descriptive mode, the two aforementioned problems can be analysed with regards to their empirical adequacy. In a normative mode, they can be analysed according to their ability to establish a justification for discounting a goodness evaluation for temporal distance.

Questions \ Modes	Descriptive	Normative
Functional Form	(i) Empirical Accuracy	(ii) Prescriptive Adequacy
Interpretation	(iii) Captures Motivation	(iv) Provides Justification

Table 4.1: Two Questions of Time Discounting and Two Modes of their Discussion

The two problems of time discounting and the two modes of their analysis yield four persistent problems of time discounting which are depicted in Table 1. We consider each of the four problems in turn.

(i) Empirical Accuracy. Firstly, consider the question of the correct functional form of time discounting in a descriptive sense. Here, we require of a time discounting function to be empirically accurate, such as accurately predicting the time discounting of agents. In this context, exponential discounting has long been defended as a sufficiently general approximation of real-world discounting behaviour, with exponential time discounting being widely employed in accounting, banking and cost-benefit analysis. However, the different variants of hyperbolic discounting functions better fit the data of experiments that have been conducted in the context of behavioural economics. As a simple illustration of this point, consider the following example: Consider an agent who is presented with two choices. Firstly, she is choosing between receiving one apple today and two apples tomorrow. Secondly, she is choosing between receiving one apple in 999 days and receiving two apples in 1,000 days. For simplicity, suppose that the agent has to make a choice, i.e. declaring her indifference is not an option. Consider that in such choices, real-world agents often choose one apple today in the first choice and two apples in 1,000 days in the second choice. Such choice behaviour is in fact predicted by hyperbolic discounting functions, whereas exponential discounting functions would imply that the agent either chooses 'symmetrically', i.e. either the one apple or the two apple in both choices. This question of greater predictive accuracy of hyperbolic discounting has enjoyed a lot of attention as of late; in particular, many different proposals for hyperbolic discounting have been made as reviewed earlier, owing to the rise of experimental and behavioural economics (Loewenstein and Prelec (1992), Loewenstein and Read (2003), Frederick *et al.* (2002)).

(ii) Prescriptive Adequacy. Secondly, consider the problem of the correct functional form of time discounting in a normative sense. Here, we require of a time discounting function to constitute the right kind of prescription of how to value future consequences. In other words, a discounting function tells us how we rationally ought to discount the future. This question has received some attention in the literature that discusses the relative merits of exponential and hyperbolic discounting: the often-rehearsed argument is that hyperbolic discounting implies preference reversal, whereas exponential discounting is dynamically consistent because it preserves the utility function (for a detailed discussion of this point, see Manzini and Mariotti (2007)). Consider again the example of the apples as introduced above. From a prescriptive point of view, dynamic consistency de-

mands that the agent chooses uniformly in the two choices, either opting for the one apple-option both times, or for the two apples-option. The choice behaviour predicted by hyperbolic discounting is dynamically inconsistent: for if the agent first chooses one apple today over two apples tomorrow, then after 999 days she will prefer to receive an apple on that day, rather than waiting for two apples on day 1,000 – but this goes against her earlier preference of rather receiving two apples in 1,000 days rather than one in 999 days. Giving up her preference for receiving one apple on day 999 for the sake of dynamic consistency does not solve the problem: on day 999, she will now have changed her preference of receiving one apple today rather than two apples tomorrow. This, in fact, is often taken as a conclusive normative argument in favour of exponential discounting, with many commentators arguing that hyperbolic discounting is only applicable as a descriptive theory (Loewenstein and Prelec (1992), Loewenstein and Read (2003), Frederick *et al.* (2002)). Note, however, that the problem of dynamic inconsistency is in fact restricted to whether discounting is in conflict with supposedly rational, or more specific, stable preferences. All it shows is that hyperbolic discounting can have the effect of undermining specific features of goodness evaluations, with that problem not arising in exponential discounting. It is not evident that this says much about the normative viability of exponential or hyperbolic discounting functions *as such.* It could simply be the case that we ought to value the future hyperbolically – and that this evaluation is in conflict with received theories of goodness evaluation. We will revisit this question later.

(iii) Captures Motivation. Thirdly, consider the question of interpreting time discounting in a descriptive sense. In this context, we require of a theory of time discounting to correctly capture the motivation for time discounting. That is to say, time discounting theories should give a substantive account of why agents engage in time discounting. In constant-rate exponential discounting theories, it is often claimed that time impatience is a psychological fact which can be captured by time preference, as reviewed earlier. In contrast, hyperbolic discounting theories maintain that preference change, attitudes towards risk and uncertainty and delay perception also play a vital role in determining how real-world agents discount for temporal distance. Since the different theories of time discounting are disparate in the conceptual content they associate with temporal distance, and in the way they interpret the discounting functions, it is not immediately obvious how the accounts can be compared to each other with regards to this question.

Little explicit attention has been given to this question in the literature which is by and large focused on the problem of predictive accuracy of time discounting functions. Indeed, somewhat anticipating the discussion in the later sections of this chapter based on the representational framework for time discounting, we can already flag the relative lack of coherent conceptual motivation for discounting the future as a key deficiency of present theories of time discounting.

(iv) Provides Justification. Fourthly, consider the question of interpreting time discounting in a normative context. Here, we require of a theory of time discounting to provide sufficient conceptual motivation to justify time discounting. That is to say, there has to be a reason as to why initial, timeless goodness evaluations of intertemporal prospects could (let alone should) be weighted by time discounting factors. In philosophy, it is mostly denied that time discounting of future goods is justifiable (e.g. Sidgwick (1907), Rawls (1971), Broome (1991), Broome (1999)). Prominently, Rawls (1971, 259) follows Sidgwick (1907) in saying that discounting future value as such is not permissible:

> 'The mere difference of location in time, of something's being earlier or later, is not a rational ground for having more or less regard for it.'

This assertion questions that the value-making features of a state of the world could depend on its temporal occurrence. Moreover, it also implies that evaluating goodness is separate and should be prioritised. Indeed, it is claimed that when evaluating goodness, the standard that is chosen (for instance, pleasure, wellbeing, or utility), should not be amended with other considerations. Hence, a separate reason is needed in order to justify time discounting. Furthermore, it suggests that time discounting will be in conflict with the assumed generality of the standard of evaluation that theories of goodness usually invoke. That is to say, upon committing to some variant of utilitarianism or expected utility, little if any room is left for influencing evaluations beyond those deemed to determine goodness in the given theory. This is in stark contrast to fact that the discounted utility-model is widely applied in economics. The somewhat unresolved situation is aptly summarised by Ramsey (1928), as cited in the introduction to this chapter. Little is offered by the time discounting interpretations themselves to improve on this; i.e. the conceptual motivations for time discounting do not carry with them a set of arguments for their normative appeal. However, note that the interpretations do come close to expected utility theory in different ways: the time preference interpretation simply extends by its very nature the familiar concept

of preferences, the uncertainty account also draws on the familiar concept of risk preferences or indeed probability functions and the preference change interpretation is also close to standard decision-theoretic accounts. While this discussion makes explicit what kinds of assumptions are made by the accounts, it does not suggest how to make these normatively plausible. Problem (iv) is hence a particularly difficult requirement to fulfil, given that time discounting modifies the goodness evaluations provided by sophisticated normative accounts, such as expected utility theory.

More generally, debates about the relative merits of time discounting theories tend to either focus on one of the four problems raised here, or pairs of problems, according to the questions (rows in Table 4.1) and modes (columns in Table 4.1). For instance, by adopting a descriptive mode, both the correct functional form and motivation of time discounting can be debated in an empirical sense. Likewise, a normative mode can be adopted, which leads to different concerns with regards to the functional form and interpretation of time discounting. Concerning the questions of time discounting, theoretical discussions in economics focus on establishing the correct functional form of time discounting in both a descriptive and normative sense whereas conceptual discussions in ethics and policy discourse focus on the question of interpretation. It is probably only the DU-model (i.e. constant-rate exponential discounting, with time preferences) that has been taken regularly as a prominent referent theory for all four of those questions. However, the DU-model can hardly be seen as providing a satisfactory answer to even one of those questions (Loewenstein and Read, 2003). Yet, as oldest fully formed theory of time discounting, it is a natural point of departure and comparison for more recent accounts and debates.

The four problems of time discounting raise deep foundational and methodological worries (although (i) could be discussed exclusively empirically). From a foundational perspective, debates about the merit of discounting theories should be resolved by analysing the theories according to their frameworks of representation and measurement.

4.3 Representation Theorems for Time Discounting

This section reviews the underlying frameworks of representation in existing theories of time discounting. We first highlight the crucial role of representation

theorems, before reviewing representation theorems for both exponential and hyperbolic discounting theories. In a next step, problems with regard to the representation theorems in the literature are raised. Finally, the strategy of the general framework of representation developed in the next section is motivated.

4.3.1 Representation Theorems and Measurement Theory

Representation theorems play a crucial role for basic economic concepts, such as utility. Indeed, for many contentious concepts in the foundations of economics, descriptive and normative debates can be explored systematically by going back to underlying theories of measurement and representation (Boumans, 2007). More specifically, such discussions are best framed in terms of representation theorems that underlie the different theories.

As a prominent example of such a strategy, consider how expected utility theory has been discussed recently in behavioural economics and philosophical decision theory. Indeed, its normative and empirical problems have been debated with reference to the structure of the representation of expected utility. For instance, in the wake of Allais (1953), the independence assumption in standard expected utility theories came under both descriptive and normative scrutiny: on the one hand, the new fields of experimental economics and behavioural economics conducted numerous studies on whether real-world agents conform to the conditions implied in standard expected utility theories. This has led to the development of amended expected utility theories (Starmer, 2000). On the other hand, there has also been conceptual and normative discussion about the conditions inherent in expected utility theory, with reference to how such theories capture the notion of risk and whether specific conditions, such as the independence assumption are normatively viable.

Indeed, there is a striking parallel between the recent debates about time discounting theories and the way in which standard expected utility theories have been challenged with empirical evidence. The four problems of time discounting described earlier roughly map onto similar descriptive and normative problems in expected utility theory. That is to say, the presence of a representational framework (in this case utility as a positive affine representation of rational preferences) has helped to structure the debates and disagreements in various ways: concerning (i), the empirical accuracy of expected utility theory, several candidate theories have emerged that are put forward as better fitting data obtained

in experiments and faring better in predictions (such as Kahneman and Tversky (1979)). Concerning (ii), the question of prescriptive accuracy, several alternative axiomatisations have been developed that avoid specific conditions; for instance, Loomes and Sugden (1982) do not include a transitivity condition and other frameworks drop the independence condition, such as Levi (1986). On a conceptual level, as in (iii), the descriptive relevance of the presentation of a decision problem ('framing') has been extensively considered (and integrated in descriptive theories, such as by Kahneman and Tversky (1979)). Farther, mirroring (iv), the normative relevance of certain conceptual limitations of standard expected utility theory have been recognised by providing representations that allow for phenomena such as taste change (Dietrich and List, 2009) and investigating dropping the completeness assumption, which has been questioned as conceptually implausible (on both a descriptive and normative level, e.g. Bradley (2009c)). This parallel between foundational problems in time discounting and expected utility theory suggests that representational frameworks are very useful for structuring the discussion of such questions.

The Representational Theory of Measurement

More formally, representation theorems make explicit the construction principles of functions. The concept of representation theorems as used in expected utility theory is closely related to the representational theory of measurement. Measurement theory is a natural starting point when considering how to set up a procedure that captures relevant features in a quantitative way (Boumans, 2007). Indeed, in their introduction to measurement theory, Savage and Ehrlich (1992, 2) maintain that

> 'measurement in general is taken to be the assignment of numbers [...] to entities and events to represent their properties and relations.'

Savage and Ehrlich (1992, 3) admit that this general characterisation of measurement is already quite close to the representational theory of measurement. The latter has been developed as a formal and abstract approach to measurement, generalising the approach to measurement that originated with physical measurement. In the latter, the central idea is that (physical) quantities are assigned numbers. This concept is captured in the following assertion of Russell (1903, 176):

> 'Measurement of magnitudes is, in its most general sense, any method by which a unique and reciprocal correspondence is established between all or some of the magnitudes of a kind and all or some of the numbers, integral, rational, or real, as the case may be ... In this general sense, measurement demands some one-one relation between the numbers and magnitudes in question – a relation which may be direct or indirect, important or trivial, according to circumstances.'

The representational theory of measurement has taken a more abstract stance, substituting the idea of physical quantity or magnitude with properties or features of objects or with relations between such properties or features (Swistak, 1990). Swistak (1990, 7) also maintains that the 'representational paradigm is the fundamental notion of measurement which is in use in the contemporary theory of measurement' and ascribes the coining of the term 'representational theory of measurement' to Adams (1966). The authoritative statement of the representational theory of measurement can be found in the three monographs Krantz *et al.* (1971), Suppes *et al.* (1971) and Luce *et al.* (1971). In their characterisation, a representational measurement procedure allows one to make two formal statements,

> 'a representation theorem, which asserts the existence of a homomorphism ϕ into a particular numerical relational structure, and a uniqueness theorem, which sets forth the permissible transformations $\phi \mapsto \phi'$ that also yield homomorphisms into the same numerical structure. A measurement procedure corresponds in the construction of a ϕ in the representation theorem.' (Krantz *et al.* (1971, 12).

Accordingly, *representation theorems* establish homomorphisms between empirical and numerical structures that allow to characterise properties of numerical assignment. For this, we assume an empirical relation R on a set of objects A and a numerical relation S on $\mathbb{R}$. A homomorphism is established by a function that assigns real numbers to elements in A in a way that numerically captures their empirical relation. More formally,

> '... if $\langle A, R_1, \ldots, R_m \rangle$ is an empirical relational structure and $\langle \mathbb{R}, S_1, \ldots, S_m \rangle$ is a numerical relational structure, a real valued function ϕ on A is a *homomorphism* if it takes each R_i into S_i, $i = 1, \ldots, m$.' (Krantz *et al.*, 1971, 8ff.)

Such homomorphisms can be characterised formally to render explicit what kinds of transformations are possible which is captured by the concept of scales:

> 'A homomorphism into the real numbers is often referred to as a scale in the psychological measurement literature. From this standpoint measurement may be regarded as the construction of homomorphisms (scales) from empirical relational structures of interest into numerical relational structures that are useful.' (Krantz *et al.*, 1971, 9)

The exact characterisation of what kind of scale a given measurement procedure yields is given by *uniqueness theorems* which specify the permissible transformations of the numbers. More formally, uniqueness theorems assert that

> '... a transformation $\phi \mapsto \phi'$ is permissible if and only if ϕ and ϕ' are both homomorphisms of $\langle A, R_1, \ldots, R_m \rangle$ into the *same* numerical structure $\langle \mathbb{R}, S_1, \ldots, S_m \rangle$.' (Krantz *et al.*, 1971, 12)

Following Steven (1946), a distinction is usually made between nominal, ordinal, interval and ratio scales. Nominal scales allow only for one-to-one transformations. Ordinal scales allow monotonic increasing transformations of the form $\phi \mapsto f(\phi)$. Interval scales allow for affine transformations of the form $\phi \mapsto \alpha\phi + \beta, \alpha > 0$. Ratio scales allow for multiplicative transformation of the form $\phi \mapsto \alpha\phi, \alpha > 0$. The representation developed here will make use of interval and ratio scales.

Since the representational framework of measurement is very general, particular frameworks have emerged, such as extensive, conjoint, bisection and difference measurement (reviewed in Suppes (2002, 63ff.)). Representations in extensive measurement specify procedures that make use of the addition of magnitudes, such as in measuring physical magnitudes of mass and length. Bisection measurement gives representations by using the operation of identifying a midpoint in an interval. Conjoint measurement representations allow the combinations of magnitudes or properties, such as when measuring the intensity and frequency of a phenomenon. In difference measurement, representations capture the intensity of a particular property or relation. Variants of difference measurement have been used for some representations of time discounting, such as by Fishburn and Rubinstein (1982) and Manzini and Mariotti (2007). (So-called absolute-difference structures will be used in the general representation developed in the next section.)

The relevance of the representational theory of measurement for developing time discounting functions is also exemplified by the fact that the development of specific formal frameworks in the representational theory of measurement is intrinsically linked with the goal of formalising measurement of psychological quantities such as preferences, emotions, beliefs etc. For an overview of the historical developments, see the references in Krantz *et al.* (1971, 9). More specifically, consider the characterisation of utility in the context of representational theory of measurement, such as in Suppes and Winet (1955) or the measurement-theoretic formulation of Ramsey's representation theorem in Bradley (2004).

This brief review of measurement theory highlights the fact that not only are representation theorems useful for greater formal and conceptual clarity in discussing scientific concepts, but that they also make explicit what kind of assumptions need to be endorsed in order to quantify qualitative properties of objects. For instance, in the context of utility, representation theorems allow us to define choice-theoretic exercises by which preferences of agents can be elicited. If those choice-theoretic exercises can be characterised by a specific set of axioms, then the preferences of agents can be represented by a utility function (unique up to choice of scale).

In the context of time discounting, the representational theory of measurement reminds us that in order to establish a function, the underlying empirical structure needs to be well-defined and -described in order to motivate the assignment of numbers. This suggests that the focus on the specific shape of discounting functions in the literature, and the relative neglect of the conceptual motivations for time discounting is problematic: the latter is much-needed in order to explain the objects in both the empirical and numerical structures that underlie representation theorems. Moreover, note that most time discounting functions reviewed earlier are considerably more complicated than utility functions. In particular, time discounting functions exhibit regularity properties (such as a constant or declining rate) with regards to a set of externally given time points, which is a formal requirement of a kind that we do not find in utility functions. This suggests that more assumptions will be required in order to establish representation theorems for time discounting than those for standard expected utility frameworks. In the following, we review representation theorems for time discounting, before raising problems concerning their comparability and the types of domains they assume.

4.3.2 Representations of Time Discounting

This section gives a brief overview of the underlying frameworks of representation in existing time discounting theories. Exponential discounting is by far the most prominent theory of discounting. As already mentioned, the canonical interpretation of time discounting in economics depicts the discount rate r in the exponential discounting function as a representation of time preference. At the heart of the time preference interpretation of time discounting lies the idea that time impatience plays a major role in intertemporal decisions. That is to say, it is assumed that agents have a preference for the present, and a preference for earlier rewards over later ones. Such preferences are supposed to be captured by the discount rate r which is then used to obtain discount factors.

It is worth noting that assuming time preference is a significant departure from and addition to standard expected utility theory. Agents as commonly modelled in decision theory and microeconomics have complete and transitive preferences over a set of prospects from which an additive utility function is derived. However, in addition to those preferences, the concept of *pure and positive time preference* is introduced to derive exponential discounting. That is, the concept of time preference introduces a new and additional type of preference. It can be further described as a structural preference as it is supposed to hold over the temporal dimension of all prospects. In addition, the fact that those time preferences are 'pure and positive' implies that an agent has a preference for utility at earlier points in time over later points in time, in addition to his preferences over prospects. More specifically, *pure* time preference refers to the fact that the betterness which is expressed by time preference is associated with distance in time. *Positive* time preference refers to the fact that time preferences capture the idea that 'earlier' is better than 'later', with the present being the earliest point in time considered such that positive outcomes at it are preferred to all later points in time.

Pure and positive time preferences are one of the key assumptions in Samuelson (1937), which has served as the standard derivation of time discounting factors in economics. Samuelson (1937) made the following assumptions in his discounted utility (DU)-model: (1) positive time preference that captures time impatience (hence, $\delta < 1$), (2) constant rate r, reflecting stable time preferences, (3) stable preferences, i.e. stationary instantaneous utility, (4) separability of utility, (5) independence of consumption, i.e. outcomes experienced at one time do not have

effects on the experience of outcomes at other times, (6) independence of discounting, i.e. the constant rate r applies to all possible outcomes. It can be shown that given the aforementioned conditions, the discounting function has to be exponential. Using a similar set of assumptions, Koopmans (1960) provided axiomatic foundations for exponential time discounting, and likewise did Lancaster (1963) and Fishburn and Rubinstein (1982). The structure shared by these axiomatisations is to postulate that agents have pure and positive time preferences and then develop conditions on those preferences that jointly capture time impatience and preserve the utility function from standard expected utility theory. The conditions used in these derivation suggest that the rate of time preference r is constant for all periods.

More formally, in the DU-model and its variants, the new concept of time preference is introduced to deal with the subjective evaluation of intertemporal prospects. Instead of introducing a valuation on some set of outcomes X which gives a utility function, time preferences compare prospects that are combinations of outcomes and times, i.e. the domain of preference becomes $X \times T$, and time preferences $\succsim_{TP}$ over this domain can be numerically represented as *discounted utility*. Fishburn and Rubinstein (1982) provide an axiomatisation of pure and positive time preference in difference structures that gives exponential discounting in similar spirit to standard representations by Samuelson (1937), Koopmans (1960) and Lancaster (1963).

The formal framework in Fishburn and Rubinstein (1982) is based on an outcome-time structure $\langle X \times T, \succsim_{TP} \rangle$, where $\succsim_{TP}$ is a preference relation on pairs of time points and outcomes (note that the outcomes are already valued, i.e. the set of outcomes is a set of utilities). Fishburn and Rubinstein (1982, 680) use the standard conditions on rational preferences as a starting point and add further conditions to prove that the preference relation $\succsim_{TP}$ can be represented by a discounted utility function.

Outcome-time structure (Fishburn and Rubinstein, 1982, 680). X is a non-degenerate real interval, T is either a set of successive non-negative integers or an interval of non-negative numbers, and $0, 1 \in T$. For $\succsim_{TP}$ on $X \times T$, for all $x, y \in X$ and all $s, t \in T$, the following conditions hold:

Weak order. $\succsim_{TP}$ is a weak order on $X \times T$;

Monotonicity. If $x > y$ then $(x, t) > (y, t)$;

Time impatience. If $s < t$ then

(i) if $x > 0$ then $(x, s) \succ_{TP} (x, t)$,

(ii) if $x = 0$ then $(x, s) \sim_{TP} (x, t)$, and

(iii) if $x < 0$ then $(x, s) \prec_{TP} (x, t)$;

Continuity. $\{(x, t) : (x, t) \succsim_{TP} (y, s)\}$ and $\{(x, t) : (y, s) \succsim_{TP} (x, t)\}$ are closed in the product topology on $X \times T$;

Stationarity. If $(x, t) \sim_{TP} (y, t + \mu)$ then $(x, s) \sim_{TP} (y, s + \mu)$.

The key conceptual assumptions in this framework are the conditions of time impatience and stationarity. The time impatience condition states that agents prefer to receive positive utility earlier and negative utility later (Fishburn and Rubinstein, 1982, 680). The stationarity condition asserts that indifference between two time-dependent outcomes depends only on the difference (μ) between the times and not on the actual time points $s, t \in T$ (Fishburn and Rubinstein, 1982, 681). Together, these constraints on the outcome-time structure $X \times T$ assure that time preferences can be represented by a discounted utility function.

Theorem (Fishburn and Rubinstein, 1982, 682). Suppose time preferences $\succsim_{TP}$ on an outcome-time structure $X \times T$ that satisfy the conditions of weak order, monotonicity, continuity, time impatience, and stationarity. Then, given any $0 < \delta < 1$, there is a continuous, increasing real-valued function u on X such that:

(i) for all $(x, t), (y, s) \in X \times T$, $(x, t) \succsim_{TP} (y, s)$ iff $\delta^t u(x) \geq \delta^s u(y)$,

(ii) for all $x \in X$, $u(x)$ is zero if $x = 0$, positive if $x > 0$, and negative if $x < 0$,

(iii) if T is an interval then u is unique (given δ) up to multiplication by positive constants on $\{x \in X : x > 0\}$ and on $\{x \in X : x < 0\}$.

Accordingly, upon assuming an outcome-time structure, time preferences are representable by a discounted utility function. From representations like the above also follows that the discounting function has to be an exponential one. More specifically, in the above theorem, the constant $0 < \delta < 1$ is the discounting factor, and it is immediately obvious that δ^t is equivalent to exponential discounting. The above representation also shows that time discounting is by no means fully determined: 'One may fix the discount factor δ arbitrarily to represent a given

preference relation that satisfies the axioms, provided the utility function u is calibrated accordingly.' (Manzini and Mariotti, 2007, 4). More precisely, we can take any two discounting factors α and β and find two utility functions such that (u, α) preferences are identical to (v, β) preferences, if they are in the same type of representation.[3]

The property of the non-unique discounting factor in the above representation is different from derivations of exponential discounting that follow Koopmans (1960), where the discounting factor is both constant and unique in the representation. This is due to the different domains which underlie the two representations, in particular the kinds of outcomes under consideration: in Koopmans (1960), time preferences are defined over consumption streams, whereas the representation in Fishburn and Rubinstein (1982) defines time preferences over single, timed outcomes. Those differences between the frameworks are not crucial to the goal of the analysis pursued here, as we focus on the more general question of how to evaluate intertemporality correctly, and how the time preference approaches attempt those evaluations. Fishburn and Rubinstein (1982) has been chosen here because it allows us to consider non-constant discounting factors by weakening the stationarity assumption, which will be discussed below.

However, the non-unique discounting factor in the above representation does highlight the fact that both an evaluation of time and an evaluation of goodness can have an influence on intertemporality in the time preference framework. The often debated problem of the 'choice of the correct discount rate' in public policy and environmental ethics debates can thus not necessarily be debated without making assumptions about goodness preferences in time preference frameworks. While the above framework makes this problem more explicit than others, most of the time preference frameworks presuppose a method of determining a unique utility function, for instance by defining time preferences over consumption streams. This renders axioms that are similar to the ones in the above representation even stronger.

On a more general level, the fundamental observation here is that time preference theories interlink time impatient attitudes with standard preferences. On

[3] As an illustration of this property, consider the following example: suppose $X = T = [0, 1]$ and u is the unique u (by (iii) in the above theorem) that satisfies the representation when $u(1) = 1$. Then, u and v are related as follows: $v(x) = [u(x)]^k$, where $k = \frac{\log \beta}{\log \alpha}$ and α, β are the respective discount factors for u and v. For $\alpha = \frac{1}{2}$ and $\beta = \frac{2}{3}$, we have $k = \frac{\log \frac{2}{3}}{\log \frac{1}{2}} = \frac{\log 2 - \log 3}{-\log 2} = \frac{\log 3 - \log 2}{\log 2}$.

the one hand, this makes available some motivations for the constraints on time preferences, such as weak order. On the other hand, the additional conditions that lead to the representation of discounted utility with an exponential discounting function are introduced without explicit further motivation. For instance, the stationarity assumption is difficult to motivate. The latter is acknowledged by the proponents in the field; indeed, Fishburn and Rubinstein (1982, 681) admit that they 'know of no persuasive argument for stationarity as a psychologically viable assumption.' While it is possible to replace stationarity with various types of separability assumptions, any time preference framework with exponential discounting as target representation will have to endorse some conditions that affect utility that go beyond those in standard representations of utility.

If the goal of a constant discounting factor is given up, weaker assumptions that stationarity can be endorsed. For instance, Fishburn and Rubinstein (1982) go on to show how replacing stationarity with an assumption of separability can yield discounting in which the discounting factor is not constant.

Thomsen separability. If $(x,t) \sim (y,s)$ and $(y,r) \sim (z,t)$ then $(x,r) \sim (z,s)$, for all $x,y,z \in X$ and all $r,s,t \in T$.

The above separability condition is weaker than stationarity and yields a representation of discounted utility in which the discounting factor does not have to be constant (Fishburn and Rubinstein, 1982, 683). The latter is indeed compatible with many variants of hyperbolic discounting in which the discount rate is declining (Manzini and Mariotti, 2007). Such variants of hyperbolic discounting theories have been derived by, for instance, Laibson (1997)), Ok and Masatlioglu (2007) and Halevy (2008). Most of these derivations, however, have not proposed entirely new frameworks of representation. Rather, they have suggested different shapes of the utility function, that is, they have proposed to amend standard DU-frameworks as the one above, introducing parameters that capture present bias, which are then motivated by diminishing impatience, the influence of time on attitudes towards risk and uncertainty, and preference change.

Exceptions to this strategy are the recent representational framework by Ok and Masatlioglu (2007) and the proposal by Scholten and Read (2006). Ok and Masatlioglu (2007) keep the standard time preference framework, but at the same time a separate, 'relative' discounting function is derived from a time-domain T, on which time intervals are compared. The relative discounting function captures

the characteristic present bias of hyperbolic discounting in the following way: time intervals in the near future are evaluated as more significant than those in the far future, yielding a relative discounting factor that declines more drastically in the near future. Scholten and Read (2006) provide a closely related account, but focus on explaining empirical evidence, rather than providing a representation. More generally, on those accounts, in addition to an outcome-time structure, a further evaluation of time intervals is assumed in order to represent hyperbolic discounting. Yet, since those frameworks additionally assume an outcome-time domain, the problems of the time preference representations apply to them as well. We will now turn to a critical review of those strategies, asking whether they provide adequate frameworks to discuss the four problems of time discounting posed in Section 4.2.

4.3.3 Problems of Time Discounting Representations

This section discusses problems with the time discounting representations introduced, focusing on the entanglement of time and value in those frameworks. Revisiting the four problems of time discounting, it can be shown that for each of those, the close interlinks between time and value in the time discounting representations are worrying. Finally, we suggest that the target of any time discounting representation, namely some variant of a decreasing function from time points to a real interval, puts severe constraints on the kind of intertemporal phenomena that can be captured by time discounting functions.

In what sense are matters of time and value interlinked in the above representations? The obvious starting point is to consider the domains and main objects in the representations. Here, both the time-outcome domain and time preferences concern time as well as value. However, it is far from obvious that this is the best way to evaluate intertemporal prospects. As briefly reviewed in Chapter 2, the value dimension of prospects can be covered by utility theories, as they provide sophisticated and well-founded tools for evaluating the goodness of prospects. The additional challenge of evaluating *intertemporal* prospects resides in the open question of how to accommodate the time dimension in procedures of evaluation. Most theories have taken the approach to model evaluations of such prospects as ones on outcome-date pairs. This has to do with the fact that there are many well-founded and successful theories for goodness evaluations of prospects. A natural consideration is to amend those already well-founded theories in

order to extend their scope. Indeed, the latter point about the well-foundedness and well-entrenchedness of theories of goodness evaluations can be taken to be at the heart of this strategy of amending the goodness evaluation with some time-related feature (such as the integration of time preferences into preference theories of goodness evaluations). The hope behind such strategies is that the normative and descriptive force of the goodness evaluation remains intact, providing the resources for the four tasks of the interpretation, while the scope of evaluation is widened, now also including time features of prospects. Yet, with ever more sophisticated attempts to determine descriptively and normatively valid ways to perform intertemporal goodness evaluations, the formal and conceptual problems in such a strategy have become more and more intractable. This is particularly pressing since time discounting lies at the heart of crucial debates about public policy, such as how to deal with climate change.[4]

In contrast, it is also possible to develop a separate evaluation of the time dimension by giving an account of how we should understand and evaluate intertemporality, and then combine such an evaluation with a goodness evaluation. Before outlining such a strategy in the next section, we will show in greater detail how the present strategy of evaluating time-outcome pairs via time preferences faces obstacles in answering the four problems of time discounting.

Functional Form

Consider firstly the question of the correct time discounting function. The framework reviewed in the previous section lends itself to comparisons of assumptions that lead to different functions, with stationarity giving exponential discounting and weaker separability assumptions being compatible with hyperbolic discounting. That is, the representation allows us to discuss the kinds of assumptions that produce either one of the standard shapes of the discounting functions, hyperbolic and exponential, in the same framework. On the downside, since the framework is a based on the idea of a utility representation of preferences over

[4]A somewhat separate, but no less pressing question is how uncertainty influences time discounting and intertemporal valuation. In the above outline of the formal elements of a discounted value representations, problems of risk and uncertainty have been assumed away. In standard frameworks, all aspects of risk and uncertainty are integrated in to a representation of *expected* value. Yet, on some accounts, time discounting is motivated by considerations of uncertainty about the future (as mentioned in Section 4.2.3). If we consider *expected* value, then delineating the risk and uncertainty that is captured by a probability function from time-related risk and uncertainty poses even further foundational challenges.

outcomes, discussing the different methods of time discounting cannot be separated from discussions about utility. This becomes relevant when considering the correct functional form descriptively as well as normatively.

In a descriptive sense, the correct functional form of time discounting should be based on empirically plausible axioms. The empirical plausibility of axioms on (goodness) preference is well-researched in the behavioural economics literature, and has led to the development of decision theories that drop some conditions that have been shown to be violated (such as transitivity, and independence). Formulating axioms on the evaluation of intertemporal prospects carries over the issues of how empirically plausible those axioms are, and given the intertemporal context, their plausibility might change. Furthermore, while the separability conditions that allow for hyperbolic discounting are weaker than stationarity, they too are not seen as very plausible (Fishburn and Rubinstein, 1982, 687). Moreover, note that such weaker conditions are only compatible with hyperbolic discounting; that is, they do not directly imply such a functional form. For this, further assumptions are made about the behaviour of the utility function which are rarely axiomatised (Manzini and Mariotti, 2007). More worryingly, many other descriptive problems with expected utility, such as those related to risk attitudes, framing effects, and preference change also need to be accommodated when investigating the evaluation of intertemporal prospects in a descriptive sense. The problem of preference change, closely related to intertemporal decisions, is especially difficult in this regard: take a decision-maker who chooses a small and early reward over a high and late reward. Does this decision-maker exhibit time preferences that suggest hyperbolic discounting or did the decision-maker undergo a momentary preference change? Such questions are hard to characterise on an outcome-time domain exactly because goodness evaluations and time evaluations are intertwined.

In a normative sense, the correct functional form of time discounting should be based on normatively plausible axioms. Philosophical decision theorists have investigated the normative plausibility of axioms on (goodness) preferences. Yet, it is unclear whether those can be transformed to an intertemporal context. Transitivity is a normatively plausible requirement on preference, but when applied to time preference, it is a much stronger assumption. There are good reasons for intransitivity of preference in an intertemporal context, for instance, when an individual experiences changes in her personality and hence her tastes. Yet, con-

ditions on time preference do not lend themselves to descriptions of what kinds of effects of intertemporality one might want to endorse. More generally, the main deficiency of representations in outcome-time structures is that the normativity of (goodness) preference get in the way of considering what the normatively correct evaluation of time distance is. In particular, it seems that variants of discounted utility live on the 'borrowed normativity' of the standard axioms on preference, which are – given the additional assumptions such as time preference and stationarity – potentially undermined by transforming them to an outcome-time domain.

Conceptual Motivation

Consider secondly the question of the conceptual motivation for time discounting. Again, the standard frameworks make it impossible to distinguish between statements that are made about time and their influence on the evaluation of intertemporal prospects on the one hand, and statements that are made about goodness evaluations of such prospects. The problems this entails for the descriptive and normative discussion of the conceptual motivation of time discounting are similar to the ones discussed concerning the functional form. Descriptively, this feature makes it hard to investigate the motivations of decision-makers for discounting. In particular, it is unclear how to integrate the diverse motivations for time discounting into a time preference framework. Normatively, it is unclear what kind of appeal an evaluation of intertemporality has: what is the normative standing of time impatience, delay perception, preference change, etc.? This question gets swept under the carpet by the concept of time preference, which seems to be taken as an innocuous modification of the normatively sound concept of (goodness) preference.

More generally, this discussion suggests that concerning all four problems of time discounting, the integrated evaluation of intertemporal features and goodness features hides contentious issues in the evaluation of intertemporal prospects. To be sure, the overall goal is to evaluating intertemporal prospects in terms of goodness. Yet, in order to do this, the influence of intertemporality has to be captured in a precise way, indeed, as precise as goodness evaluations. The problem of entanglement of time and value creates the need for a framework that evaluates temporal features separately.

Towards General Foundations

Considering the review of representation theorems in measurement theory, it becomes clear that an unambiguous description of the empirical structure is needed. Moreover, given that time discounting functions provide a very specific way to evaluate intertemporality, in that they in general assume a regular behaviour of the weights and sometimes impose further restrictions, it is important to ask what kind of intertemporal phenomena they are capable of capturing. Since time discounting functions are decreasing, the phenomena that such functions could be representing need to behave in a regular fashion. Not all problems that arise in intertemporal decisions lend themselves to motivating an empirical structure that can be represented by a time discounting function. What is required in this context are phenomena that behave in a regular fashion according to a time-index. That is to say, we need to endorse the assumption that the conceptual motivations offered for time discounting, such as time impatience, attitudes to risk and uncertainty, and preference change indeed behave in a regular way. This creates the need for a framework which can facilitate a comparison of the kinds of qualitative properties that are represented numerically by a discounting factor in the different theories.

4.4 General Foundations of Time Discounting

This section provides general measurement-theoretic foundations of time discounting. We proceed in four steps: firstly, we motivate and explain the strategy of representation as one that evaluates time distance features of prospects separately from goodness evaluations. Secondly, a numerical representation of evaluations of 'time distance' is developed. Thirdly, we discuss how standard accounts of time discounting interpret time distance, that is, what kind of time distance features they take to motivate time discounting. Fourthly, we give time discounting functions in terms of time distance and show how further restrictions yield exponential and hyperbolic discounting functions. Finally, we show how such discounting functions can be combined with goodness evaluations so as to yield discounted value.

4.4.1 Introduction

As stated above, one of the very aims of the representation developed here is to provide a general framework for the clarification of existing theories of time discounting. The most important feature of the representation is that we initially consider the time dimension and the value dimension *separately*. This is not to argue that they are or should be in some sense separate – on the contrary, it seems natural that there are various interrelations between those two dimensions. However, in order to assess how time discounting theories can achieve their goal of establishing that goodness evaluations can be weighted with time discounting factors, a separate reconstruction is in order that will make transparent what kind of assumptions about time, value, and their interrelations have to be endorsed in order to establish time discounting.

Recall that theories of time discounting offer different conceptual motivations for time discounting. The representation developed here is conceptually neutral and can be interpreted with any of the conceptual motivations discussed (and other ones). As a placeholder for the precise conceptual motivation for time discounting we will henceforth use the term 'time distance'. That is, we will show how time discounting functions can be motivated by an evaluation of salient features of time distance. To give an example, time preference theories maintain that time discounting functions capture impatient attitudes of agents to time distance. Here, we give general conditions of how evaluations of time distance can motivate time discounting.

More precisely, we embark from the supposition that a goodness evaluation of prospects can be given by a variant of decision theory and that if we are to evaluate intertemporal prospects, we also require an evaluation of the time distance features of those prospects that is then combined with the goodness evaluation. Hence, a *separate* formal procedure will be developed that enables us to evaluate the time distance features of intertemporal prospects. Once such a formal procedure is in place, it can be supplemented with a substantial interpretation that motivates the combination of the goodness evaluation with the time distance feature evaluation. The main task of the whole of Section 4.4 is to establish a detailed procedure that gives an evaluation of time distance features of prospects based on a measurement-theoretic framework. Matters of interpretation and comparison of existing theories will only alluded to briefly, and taken up in more detail in Section 4.5 and 4.6.

As the target of the following representation, consider a general definition of a time discounting function that assigns numerical values to points in time. Note that henceforth, decreasing means strictly decreasing.

Definition 1 (Time discounting function). *A* time discounting function D *is a decreasing mapping* $D : T \to (0, 1]$, *from a set of time points* T*containing* 0 *to the real interval* $(0, 1]$, *such that* $D(0) = 1$.

Accordingly, a time discounting function assigns numerical values to time points. This definition could be generalised further, but as it stands includes already most common time functions proposed in the literature, as reviewed earlier. Definition 1 fixes the target of the representation, that is, we investigate what kind of measurement-theoretic assumptions are required such that D is well-founded. More specifically, we ask: what kind of qualitative phenomena that are associated with time points are expressed by the numerical values given by this function? In order to answer this question without pre-commitment to a specific conceptual motivation for time discounting at this point, we will make the general assumption that the numerical values given by a well-founded discounting function are the result of an evaluation of salient time distance features of a set of events.

We now introduce some general notation and primitives of the representation framework. Firstly, let T be a set of externally given time points. Formally, we assume that $T \subseteq [0, \infty)$ with $0 \in T$. The point 0 represents the present and T does not contain past time points. T could be discrete (e.g., $T = \{0, 1, 2\}$) or continuous (e.g. $T = [0, 100]$) and might even have infinite horizon (e.g. $T = \{0, 1, 2, ...\}$ or $T = [0, \infty)$). We will sometimes refer to T as representing *clock-time.*

Secondly, let Q be a set of events, and let $p \in Q$ denoting the present event. We interpret sets of events as prospects, denoted $A \subset Q$. Let there be a function $\tau : Q \to T$ mapping each event to the clock-time at which it occurs. We call τ the *clock-time function.* By assumption, $\tau(p) = 0$ (i.e., the clock is set such that the present event happens at clock-time 0). We also assume that for each clock-time $t \in T$ there is at least one event $q \in Q$ with $\tau(q) = t$. That is, at each clock-time at least something happens. Having established clock-time and events, we also need an evaluation of the salient features of time distance of those events.

Thirdly, let there be a function $\varphi : Q \to I$ mapping each event to a numerical evaluation of its time distance features. By assumption, for each $i \in I$ there is at least one event $q \in Q$ with $\varphi(q) = i$. Intuitively, φ gives an evaluation of salient

features of temporal distance. We call φ the *time distance function* and give an axiomatic derivation of it later.

The goal of the following representation is to show that it is possible to construct a well-founded time discounting function by using the time distance evaluation I. The representation proceeds in four steps: firstly, we present an axiomatic derivation of a time distance function φ, in a framework of ordinal distance measurement. Secondly, we discuss how this framework can be interpreted by existing conceptions of time discounting. Thirdly, we construct a composite function which maps clock-time points T to time distance evaluations I given by φ, and the latter into a real-valued interval $(0, 1]$, such that this composite function is a time discounting function as stated in Definition 1. Finally, after giving conditions for obtaining exponential and hyperbolic discounting in this framework, we combine such a time discounting function with an evaluation of the goodness of consequences in Q to obtain discounted value.

4.4.2 Representing Time Distance

This section develops a numerical representation of the ordinal concept of 'time distance'; that is, we derive the time distance function $\varphi : Q \to I$ in a measurement-theoretic framework. As mentioned earlier, the concept of time distance is introduced as a formal notion, and is intended as a placeholder for specific interpretations which will be discussed in the next section.

The representation is closely related to a formal framework of measurement and representation developed in Krantz *et al.* (1971). Indeed, we will employ a modified variant of *difference measurement*, given in Krantz *et al.* (1971, 170ff.), to obtain a formal characterisation of a general evaluation of time distance features of intertemporal consequences. In Krantz *et al.* (1971, 170ff.), so-called 'absolute-difference structures' are introduced to measure differences along a single dimension between *pairs* of elements in a set. The following structure and representation are close corollaries of this result. Firstly, consider ordinal distance between pairs of elements in a set.

Definition 2 (Ordinal distance). *An* ordinal distance *is a binary relation* $\succcurlyeq$ *on* $Q \times Q$.

Henceforth, we write the pair $(q, r) \in Q \times Q$ as $qr \in Q \times Q$. Accordingly, ordinal distance compares distances between pairs of elements $qr \in Q \times Q$. For example,

$qr \succcurlyeq st$ means that the distance between q and r is at least as large as the distance between s and t. Note that in order to employ such a concept of ordinal distance, a single dimension of comparison needs to be specified. For instance, the elements could be compared according to their distance in their sweetness, or loudness, or how an agent perceives of their temporal distance. Consider the following axioms on ordinal distance.

Definition 3 (Ordinal distance structure). *Suppose a set Q with at least two elements and ordinal distance $\succcurlyeq$. The pair $\langle Q \times Q, \succcurlyeq \rangle$ is an* ordinal distance structure *iff, for all $q, r, s, t, q', r', s', t' \in Q$, and all sequences $q_1, q_2, \ldots, q_i, \ldots \in Q$ the following axioms hold:*

1. *Weak ordering.*
   - *(i) Either $qr \succcurlyeq st$ or $st \succcurlyeq qr$.*
   - *(ii) If $qr \succcurlyeq st$, $st \succcurlyeq q'r'$ then $qr \succcurlyeq q'r'$.*
2. *Weak symmetry.*
   - *(i) $qr \sim rq \succcurlyeq qq \sim rr$.*
   - *(ii) If $qr \sim qq$, then $qs \sim rs$.*
3. *Well-Behavedness. If $rs \succ rr$, $qs \succcurlyeq qr, rs$ and $rt \succcurlyeq rs, st$, then $qt \succcurlyeq qs, rt$.*
4. *Weak Monotonicity. Suppose that $qs \succcurlyeq qr, rs$. If $qr \succcurlyeq q'r'$ and $rs \succcurlyeq r's'$, then $qs \succcurlyeq q's'$; moreover if either $qr \succ q'r'$ or $rs \succ r's'$, then $qs \succ q's'$.*
5. *Solvability. If $qr \succcurlyeq st$, then there exists $t' \in Q$, such that $qr \succcurlyeq t'r$ and $qt' \sim st$.*
6. *Archimedean property. If $q_1, q_2, \ldots, q_i, \ldots$ is a strictly bounded standard sequence (i.e., there exist $t', t'' \in Q$, such that for all $i = 1, 2, \ldots$, $t't'' \succ q_{i+1}q_1 \succcurlyeq q_iq_1$ and $q_{i+1}q_i \sim q_2q_1 \succ q_1q_1$), then the sequence is finite.*

To illustrate these conditions, recall that Q is a set of events, and $A \subset Q$ can be prospects. For example, take the prospect of a dinner $A = \{q, r, s, t\}$, where q = starter, r = main, s = dessert and t = coffee. According to the above definition, pairs of elements in this prospect can be compared according to their distance on a single dimension. For instance, the events of the dinner prospect can be compared with regards to their sweetness. In order to do so, the pairs of

elements are ordered according to $\succcurlyeq$. Take the pairs qr (starter, main) and st (dessert, coffee). If the distance in sweetness between main and starter is smaller than that between coffee and dessert, then $st \succ qr$. The symmetry condition on this ordering states that the distance between pairs of elements is independent of their ordering, i.e. when comparing the absolute distance in sweetness between starter and main to another pair of events, it does not matter whether we write qr or rq. Indeed, the two most important properties of the ordinal distance structure from an interpretative point of view are the facts that the relation $\succcurlyeq$ orders pairs of elements (and not single elements) according to Condition 1 and that the ordering relation is symmetric according to Condition 2.

Conditions 3-6 in the above definition ensure richness of the ordering $\succcurlyeq$. They can be explained more intuitively by introducing the notion of *betweenness* that can hold for single elements. Consider three elements $q, r, s \in Q$. Then, r is between q and s iff $qs \succcurlyeq qr, rs$. This is denoted $q|r|s$. The notion of betweenness allows us to understand the relation $\succcurlyeq$ which orders pairs $qr \in Q \times Q$ in terms of *single* elements and their position relative to each other. From Conditions 1 and 2, it follows that for any element $q, r, s \in Q$, at least one of the betweenness patterns $q|r|s$, $q|s|r$ or $r|q|s$ must hold and that betweenness is symmetric, i.e. $q|r|s$ iff $s|r|q$. We can now rewrite Conditions 3-6 in Definition 3 in this more intuitive terminology:

3. Well-Behavedness. If $rs \succ rr$, $q|r|s$ and $r|s|t$, then both $q|r|t$ and $q|s|t$.

4. Weak Monotonicity. If $q|r|s$, $q'|r'|s'$, and $qr \sim q'r'$, then $rs \succcurlyeq r's'$ iff $qs \succcurlyeq q's'$.

5. Solvability. If $qr \succcurlyeq st$, then there exists $t' \in Q$, such that $q|t'|r$ and $qt' \sim st$.

6. Archimedean property. If $q_1, q_2, \ldots, q_i, \ldots$ is a strictly bounded standard sequence (i.e., if $q_{i+1}|q_i|q_1$, for all $i = 1, 2, \ldots$, and successive intervals are equal and nonnull and $q_i q_1$ is strictly bounded), then the sequence is finite.

The above conditions jointly ensure that the ordinal distance structure is sufficiently rich and well-behaved in order for $\succcurlyeq$ to be representable numerically.

Ordinal distance structures as given in Definition 3 are closely related to so-called absolute-difference structures in Krantz *et al.* (1971, 170). In the latter, the following symmetry condition is used: if $q \neq r$, then $qr \sim rq \succ qq \sim rr$. However, this rules out that Q can contain distinct events which are similar under the

domain of comparison. In Definition 3, the weaker symmetry condition 2 avoids this problem. (We show how the two types of structures just mentioned relate to each other by introducing equivalence classes when proving Theorem 5. Proofs of formal statements are given in an appendix to this chapter.)

The ordinal distance structure given here can be interpreted in a variety of ways. In order to do so, one has to define the single dimension of comparison between elements in Q and then interpret the above conditions on the ordinal distance $\succcurlyeq$ between pairs. In the example of the dinner, one could for instance change the single domain of comparison to that of time distance as perceived by agents. Then, $\succcurlyeq$ reflects the ordinal distance in time between all pairs of elements of the dinner. For instance, the time distance between starter and main qr could be perceived as greater than that between dessert and coffee st, due to livelier conversation later in the evening, such that $qr \succ st$. The next section will discuss in detail how to interpret ordinal distance by the conceptions of time distance that can underly different theories of time discounting.

Note that ordinal distances do not need to correspond to any supposedly objective standard that is externally given, such as sweetness defined in terms of sugar content in food items or time as it is given by a clock. What is crucial is the fact that if one can identify one dimension on which ordinal distances between pairs of elements can be compared, and upon the comparison satisfying the conditions in Definition 3, the ordering $\succcurlyeq$ can be represented numerically.

Definition 4 (Representation). *A real-valued function φ on Q is said to* represent *ordinal distance $\succcurlyeq$ if for all $p, q, r, s \in Q$,*

$$qr \succcurlyeq st \text{ iff } |\varphi(q) - \varphi(r)| \geq |\varphi(s) - \varphi(t)|.$$

Accordingly, $\succcurlyeq$ has a numerical representation if the absolute difference between the assigned numbers of pairs of elements adequately reflects their ordinal distance.

Theorem 5 (Interval Representation of Ordinal Distance). *Suppose ordinal distance $\succcurlyeq$ satisfies the conditions of an ordinal distance structure. Then there exists a function $\varphi : Q \to \mathbb{R}$ that* represents *$\succcurlyeq$. If φ' is another function with the same property, then $\varphi' = \alpha\varphi + \beta$, where $\alpha, \beta \in \mathbb{R}, \alpha \neq 0$.*

Accordingly, it is possible to numerically represent the ordinal distance between pairs of elements in a set Q. That is, in the context of measuring ordinal distance

between consequences on a single dimension, a number $\varphi \in \mathbb{R}$ can be assigned to any event $q \in Q$ such that for any two events $qr \in Q \times Q$, the absolute difference of $\varphi(q)$ and $\varphi(r)$ adequately reflects their ordinal distance when compared to any other pair of elements $st \in Q \times Q$ and the numbers assigned to them, such that $qr \succcurlyeq st$ iff $|\varphi(q) - \varphi(r)| \geq |\varphi(s) - \varphi(t)|$.

In the context of the dinner example and the sweetness dimension of comparison, this means that all consequences in the dinner are assigned a real number $\varphi \in \mathbb{R}$ that reflects the ordering in ordinal distance of sweetness between all possible pairings. For example, take the following ordinal distance in sweetness of the dinner consequences: $st \succ sr \succ qt \sim qs \succ rt \succ rq$. According to Theorem 5, this can be represented numerically, for instance by the following assignment of numbers: $\varphi(q) = 55$, $\varphi(r) = 50$, $\varphi(s) = 105$, $\varphi(t) = 5$. This clearly satisfies Theorem 5, as $|\varphi(s) - \varphi(t)| > |\varphi(s) - \varphi(r)| > |\varphi(q) - \varphi(t)| = |\varphi(q) - \varphi(s)| > |\varphi(r) - \varphi(t)| > |\varphi(r) - \varphi(q)|$, i.e. $100 > 55 > 50 = 50 > 45$. Moreover, the assigned numbers are unique up to an affine transformation.

For some dimensions of comparison, it is possible to identify a specific element $p \in Q$ to which the above representation can be normalised. Generally, such a normalisation is permissible if p is a true zero point.

Corollary 6 (Normalisation). *Let $\succcurlyeq$ satisfy the conditions of an ordinal distance structure and be represented by σ. Then, φ is a normalisation of σ to p iff $\varphi = \sigma + \beta$ and $\varphi(p) = 0$. If φ' is another function with the same property, then $\varphi' = \alpha\varphi$, where $\alpha \in \mathbb{R}, \alpha > 0$.*

The above statement asserts that the interval scale given in Theorem 5 can be normalised to an absolute zero which gives a ratio scale on which only multiplicative transformations are allowed. That is, to continue the example of the dinner items, if there is an element that has maximal or minimal sweetness, then the numerical representation can be normalised. Suppose an agent has a double espresso with no sugar for coffee and that this is minimally sweet. Normalising according to Corollary 6 to $t \in Q$, by the transformation $\varphi = \sigma - 5$, gives $\varphi(t) = 0$ (and, accordingly, $\varphi(q) = 50$, $\varphi(r) = 45$, and $\varphi(s) = 100$).

Interpreting the above framework with other dimensions of ordinal distance, similar orderings and representations can be given. In the context of time distance, it is indeed also plausible to normalise the representation to a ratio scale, when taking the present p as an absolute zero, as it is the natural viewpoint from which prospects and courses of actions are assessed. It is also possible to

normalise to any other time point in the past or future, which is plausible when there is a specific point in time from which temporally extended prospects are analysed. For most applications and indeed for representing time discounting, normalising to the present is the most plausible option. This normalisation also makes it possible to specifically analyse time distance between p and other elements in the set Q. Notably, from the normalisation, the following statement follows immediately:

$$rp \succcurlyeq qp \text{ iff } |\varphi^*(r)| \geq |\varphi^*(q)|.$$

Accordingly, consequences $q \in Q$ can be analysed directly with regards to their time distance to the present p. This more intuitive comparison of time distance will be used to consider the interpretations of time distance that is inherent in the different theories of time discounting.

The above representation rests on Definition 3 which contains richness assumptions which may not be satisfied in modelling applications. For instance, the assumptions imply that Q is infinite and the image of φ is an interval (or ratio). Such richness assumptions are standardly included in measurement-theoretic representations and make explicit the requirements needed to construct functions with scale-properties. In the following, we use φ also in 'non-rich' cases where, for instance, Q is finite and there are only finitely many time distances. That is, $\varphi : Q \to I$ is an (onto) function from events to their time distance. From now on, we assume that $I \subseteq [0, \infty)$ with $\sigma(p) = 0$. That is, past events are not included.[5]

Before employing the function $\varphi : Q \to I$ to derive time discounting functions, we consider how the formal notion of 'time distance' that has been given measurement-theoretic foundations in this section can be interpreted with the concepts inherent in existing theories of time discounting.

4.4.3 Interpreting Time Distance

Time distance can be interpreted in a variety of ways. Firstly, it could be interpreted as equivalent to clock-time. Indeed, the six axioms on the binary relation $\succcurlyeq$ would follow immediately from the idea that time distance is given by clock-time. It is possible to understand the framework introduced here in such a naive

[5] If φ is indeed thought of as being derived from an ordinal distance structure $\langle Q \times Q, \succcurlyeq \rangle$ with $\varphi(p) = 0$, then the assumption that $I \subseteq [0, \infty)$ implies that the present event p is not strictly between any two other events (i.e., there are no events $r, s \in Q$ such that $rs \succ rp$ and $rs \succ sp$). All other events $q \in Q$ are such that $qp \succ qq$ for future events, and Q contains no past events.

sense, yet it is unclear what this would add: the real purpose of the framework is of course not to capture measurement of time as clock-time, but to develop a framework that can compare what properties different theories of time discounting take to be relevant about time distance. That is, the term 'time distance' – as employed so far – can be understood as a formal notion for which certain properties have been defined that are required to obtain a numerical representation of temporal features of prospects.

We now turn to the question of supplementing the ordinal distance structure and its representation with substantive interpretations. Intuitively, such an interpretation will specify what one takes the relevance of distance in time to be, i.e. what kind of regular phenomena one associates with the passage of time. For all interpretations, Q depicts a set of events and $\succcurlyeq$ orders pairs of those according to their time distance. However, the approaches can differ widely in how exactly that distance is interpreted and how rich the description of the events needs to be.

Time impatience. Time preference theories of discounting evaluate time distance according to the degree of impatience it induces in the agent. Indeed, at the heart of these theories lies the idea that time impatience of agents is both psychologically plausible and plays a major role in intertemporal evaluations (Frederick *et al.*, 2002). In those theories, $rp \succcurlyeq qp$ iff a higher degree of impatience is associated with r than with q, as no time impatience is associated with p. With the additional assumption that events $q \in Q$ take a positive value under a desirability evaluation, this captures time preferences as used in the representations of Samuelson (1937), Koopmans (1960), and Fishburn and Rubinstein (1982).

Delay. Delay theories of time discounting evaluate time distance according to how agents perceive the delay inherent in it. In those theories, initiated by Ainslie (1975), Ainslie (1992), Ainslie (2001) amongst others, empirical results on how agents perceive of delays are generalised and used to motivate time discounting, and Ok and Masatlioglu (2007) provide a representation theorem that includes delay perception. Hence, in this interpretation, $rp \succcurlyeq qp$ iff an agent perceives a longer delay between rp than with qp, and no delay is associated with p.

Risk and uncertainty. Risk and uncertainty theories of time discounting evaluate time distance according to the degree of fundamental risk or uncertainty it induces. Such accounts, in which time-indexed probability functions and risk evaluations motivate time discounting are discussed by Weitzman (2001), Gollier

(2002), Mas-Colell *et al.* (1995) and Halevy (2008). Hence, $rp \succcurlyeq qp$ iff more fundamental uncertainty is associated with r than with q, as no fundamental uncertainty is associated with p. Additional assumptions on how the risk and uncertainty evaluation of time distance is delineated from risk and uncertainty in goodness evaluations are needed to employ time discounting functions thus motivated.

Preference change. Preference change theories of time discounting evaluate time distance according to the degree of change in the propositional attitudes of agents. In those theories, the future goodness evaluations of agents are discounted with their diminished present credibility due to changes in preferences, as suggested by Strotz (1956), Parfit (1984) and Frederick *et al.* (2002, 389). On such accounts, $rp \succcurlyeq qp$ iff there is more preference change associated with r than with q, when compared to preferences at p. In order for this interpretation to hold, richer descriptions of consequences have to be assumed (e.g. events $q \in Q$ are interpreted as an agent-relative proposition 'Agent α eats a dessert'), such that the description specifically includes a reference to the agent whose preferences change.

In addition to those interpretations, there is a large class of time discounting theories that combine several interpretations of time distance (overviews are in Frederick *et al.* (2002), Loewenstein and Read (2003)). As presented in the overview of hyperbolic discounting, there are theories that combine considerations of time preferences, delay, risk and uncertainty and preference change.[6] This will require us to model more than one understanding of time distance in order to be able to formulate representations of time discounting functions. For this, we either have to repeat the representation according to the different interpretations or interpret $\succcurlyeq$ as capturing all of those interpretations in one ordering.

Interpreting ordinal time distance and its representation with the conceptual content from the different theories of time discounting makes transparent how those theories establish the numerical representation of a qualitative concept. Furthermore, it also makes transparent that the specific interpretation of time distance has to motivate the ordinal distance structure given in the previous section. More generally, the framework given in this paper allows us to recast the

[6]Note that there is also a rich literature on intertemporal allocation that employs a plethora of concepts to motivate time discounting, including those already mentioned, as well as market rates of return on investment and other factors related to investment. Such specifications of 'time discounting' will be even more complex to represent than the above.

conceptions of time discounting the different theories endorse in terms of their inherent interpretation of time distance. This facilitates a clarification of their assumptions. These questions will be discussed in more detail in Section 4.5 and 4.6.

4.4.4 Time Distance Discounting

This section constructs a composite function which maps clock-time points T to time distance evaluations I given by φ, and the latter into a real-valued interval $(0, 1]$, such that this composite function satisfies Definition 1. This composite function is written as $Disc \circ c$, where c is an increasing function $c : T \to I$ and $Disc : I \to (0, 1]$ is a decreasing function with $Disc(0) = 1$. Firstly, we consider $Disc$, followed by c, to finally show that $Disc \circ c$ is a time discounting function.

Recall that $\varphi : Q \to I$ is an (onto) function from events to their time distance, with $I \subseteq [0, \infty)$ and $\varphi(p) = 0$. Here, we use the function φ to construct discounting weights. From now on let $Disc : I \to (0, 1]$ be a decreasing function with $Disc(0) = 1$. Since this function uses the representation of ordinal distance between events to formulate weights $(0, 1]$, it is called the *distance discounting function.*

Proposition 7 (Distance discounting). *Suppose φ represents ordinal distance $\succcurlyeq$ on $Q \times Q$. Then, $Disc(\varphi(q)) \geq Disc(\varphi(r))$ iff $rp \succcurlyeq qp$, for all $q, r \in Q$.*

Accordingly, the distance discounting function assigns the unit weight to $\varphi(p)$, resulting in no discounting at all for this event, and assigns a number in the real interval $(0, 1)$ to all other distances such that the larger the distance, the lower the weight. On the conceptual level, any of the specific time distance interpretations introduced earlier can be used to interpret such a distance discounting function. This completes the first step of constructing the composite function $Disc \circ c$.

As alluded to before, time distance does not necessarily correspond to time as commonly understood as clock-time. This is due to the fact that not all conceptions of how we understand and measure time distance need to correspond to clock-time. For example, imagine that what is relevant about temporal features of events is how an agent subjectively perceives of their time distance to the present. It is immediately obvious that such a subjective perception of time distance does not need to correspond in any regular way to clock-time: events quite far in the future could be perceived as close, and events in the near future could

be perceived as far away. Crucially, this could even be the case if the subjective time perception of that agent is measurable by the time distance measurement procedure introduced earlier, as it has been given as independent of clock-time. For example, let $p, q, r, s \in Q$ denote events on different days. Assume that the way the agent perceives of the time distance between those events yields the assignment $\varphi(p) = 0$, $\varphi(q) = 1$, $\varphi(r) = 2$, $\varphi(s) = 3$. However, it could be the case that the clock-time function τ assigns the following values: $\tau(p) =$ today, $\tau(q) =$ yesterday, $\tau(r) =$ in two months, $\tau(s) =$ next week. While it is still possible to construct a distance discounting function *Disc* that is meaningful – indeed, it can be used to weight consequences according to an agent's subjective time perception – it surely does not satisfy Definition 1 in this case, recalling that in the general time discounting definition it is defined as a mapping from a set of clock-time points T to a real interval.

Hence, a set of requirements is needed that ensures a correspondence between the representation of time distance and clock-time. Indeed, it is a virtue of the general framework developed here that it makes explicit this crucial regularity assumption that is implicit in any time discounting theory that gives time discounting functions satisfying Definition 1. Indeed, as will be discussed in more detail in Section 4.5 and 4.6, this implicit regularity assumption severely constrains the kinds of evaluations of time distance that time discounting theories are capable of expressing. In the following, we will state assumptions about the correspondence between clock-time and time distance that are needed to construct well-founded time discounting functions that satisfy Definition 1.

Definition 8 (Correspondence between clock-time and time distance). *A correspondence between clock time and perceived time is a function $c : T \to I$ such that $\varphi = c \circ \tau$.*

Informally, such a correspondence maps every clock-time to the time distance of the events happening at that clock-time. Recall that clock-time is defined as $T \subseteq [0, \infty)$ with $0 \in T$, and that there is a clock-time function $\tau : Q \to T$, with $\tau(p) = 0$, mapping each event to the clock-time at which it occurs. Note that a correspondence satisfies $c(0) = 0$, i.e., the clock-wise present is also the present under a time distance evaluation, due to $\tau(p) = 0$ and $\varphi(p) = 0$.

Proposition 9 (Uniqueness of correspondence). *There is at most* one *correspondence between clock-time and time distance.*

By this uniqueness result, we can henceforth talk unambiguously of *the* correspondence between clock time and perceived time, if it exists.

Proposition 10 (Existence of correspondence). *There* exists *a correspondence between clock-time and time distance iff for all events* $q, r \in Q$, $\tau(q) = \tau(r) \Leftrightarrow \varphi(q) = \varphi(r)$.

This condition for existence of a correspondence states that events which are simultaneous under clock-time also have the same time distance. We will from now on assume that this condition is satisfied. Therefore, a (unique) correspondence exists, henceforth denoted by c.

Proposition 11 (Increasing correspondence). *The correspondence* c *is* increasing *iff for all events* $q, r \in Q$, $\tau(q) > \tau(r) \Leftrightarrow \varphi(q) > \varphi(r)$.

Informally, increasing correspondence asserts that events which are later by clock-time are also characterised as later by time distance. Note that, by applying this condition in both directions, it follows that for all events $q, r \in Q$, $\tau(q) = \tau(r) \Leftrightarrow \varphi(q) = \varphi(r)$, i.e. that events which are simultaneous under clock-time are also simultaneous under time distance.

Increasing correspondence between clock-time and time distance is a substantial assumption in a conceptual sense: as discussed earlier, the degree of impatience, fundamental uncertainty, preference change or delay perception that motivates time distance could be influenced by a number of other factors and, for instance, fluctuate when compared to an externally given time index. The latter could indeed rule out that time distance understood by such conceptions is a suitable motivation for time discounting. However, given an increasing correspondence between clock-time and time distance, time discounting as stated in Definition 1 can be obtained.

Theorem 12 (Time discounting and time distance). *The following are equivalent:*

(i) D *is a time discounting function.*

(ii) c *is increasing.*

(iii) For all events $q, r \in Q$, $\tau(q) > \tau(r) \Leftrightarrow \varphi(q) > \varphi(r)$.

Accordingly, a time discounting function is a composite function which maps clock-time points T to time distance evaluations I given by φ, and the latter into a real-valued interval $(0, 1]$, such that this composite function is a time discounting function as stated in Definition 1. Indeed, the distance discounting function $Disc$ generates a decreasing function $D := Disc \circ c$ from T to $(0, 1]$, such that $D(0) = 1$. In other words, taking any pair of events $qr \in Q \times Q$ and comparing their time distance with other pairs, each consequence can be assigned a number φ that indicates their time distance. Given increasing correspondence between clock-time and time distance, the image of φ (denoted I) can then used to obtain the function $Disc \circ c$ which is a time discounting function according to Definition 1.

More generally, the general representation theorem makes transparent the requirements any theory of time discounting has to fulfil in a measurement-theoretic framework: namely, a conceptual interpretation of time distance has to be given that renders plausible the representation procedure in an ordinal distance structure. Furthermore, increasing correspondence between clock-time and time distance has to hold under this interpretation. This suggests that the conceptual interpretation of time distance is subject to two rather strong regularity assumptions. Furthermore, note that fulfilling the above requirements does not directly imply descriptive or normative plausibility with regards to discounting utility in intertemporal decisions. Indeed, time distance discounting as obtained above only gives a procedure for the evaluation of the temporal features of intertemporal prospects by well-founded time-indexed weights. We will discuss this set of general requirements for time discounting theories in greater detail in Sections 4.5 and 4.6.

4.4.5 Exponential and Hyperbolic Time Discounting Functions

The construction of a well-founded discounting function in the general framework is both in line with existing representations of time discounting and permissive enough to obtain specific functional forms of time discounting. This section shows how the time discounting function given in the general framework can be further restricted so as to obtain more specific time discounting functions that have been proposed in the literature.

Formally, exponential time discounting functions and (many) hyperbolic time discounting functions are special cases of time discounting functions as defined in Definition 1. Firstly, consider exponential discounting.

Definition 13 (Exponential time discounting). *A time discounting function D is* exponential *iff there exists a $\delta \in (0,1)$ such that $D(t) = \delta^t$ for all $t \in T$.*

It is immediately obvious that Definition 13 is a special case of Definition 1. Accordingly, exponential time discounting introduces a constant discounting factor δ which is used to calculate the discounting factor for each point in time.

In order to derive such a more restrictive time discounting function in the general framework given above, additional requirements are imposed on the correspondence between clock-time and time distance. Indeed, to obtain exponential time discounting, we assume linearity of the correspondence c. Examples of such linear correspondence c are, for instance, $\varphi = \tau$, or $\varphi = 2\tau$. Consider the latter for $T = \{0,1,2\}$ and $I = \{0,2,4\}$. Here, $c : T \to I$ is given by $c(0) = 0$, $c(1) = 2$ and $c(2) = 4$, i.e. $k = 2$. A more complex example is $T = I = [0,\infty)$ and $c(t) = 3t$, for all $t \in [0,\infty)$. In other words, time distance is proportional to clock-time iff correspondence c is linear. If linear correspondence holds, exponential time discounting functions can be given in the general representational framework.

Theorem 14 (Exponential time distance discounting). *Suppose c is increasing, so that D is a time discounting function by Theorem 12. If time distance is proportional to clock-time and Disc preserves proportionality, then D is exponential.*

Accordingly, if linear correspondence ensures the proportionality between clock-time and time distance and *Disc* preserves the proportionality, D is an exponential time discounting function.

On a conceptual level, note that the assumption of linear correspondence translates into an assumption about the regularity of time distance. More specifically, in order for time distance to be proportional to clock-time, the time distance between any two events that are subsequent under (discrete) clock-time must be the same.

Naturally, time distance does not need to behave in such a regular manner so that proportionality holds. Another possibility is that time distance might be concave in clock-time, i.e. $\varphi = c \circ \tau$ for some concave function $c : T \to I$.[7] Examples of concave correspondence c include $\varphi(q) = \sqrt{\tau(q)}$ for all $t \in T$ or $T = I = [0,\infty)$ and $c(t) = \sqrt{t}$ for all $t \in T$, or $c(t) = \log(1+t)$ for all $t \in T$.

[7] That is, $c(at + (1-a)t') > ac(t) + (1-a)c(t')$ for all distinct $t, t' \in T$ and all $a \in (0,1)$ such that $at + (1-a)t' \in T$.

That is to say, time distance is concave in clock-time iff c is concave. Such an assumption can be used to characterise hyperbolic discounting.

Theorem 15 (Non-exponential time distance discounting). *Suppose c is increasing, so that D is a time discounting function by Theorem 12. If time distance is concave in clock-time, Disc preserves concavity, and T contains more than two time points (for non-triviality), then D is* non-exponential.

Accordingly, if time distance is concave in clock-time, we obtain non-exponential time discounting. This is significant as many hyperbolic discounting functions exhibit properties that can be described as concave correspondence. Due to the plethora of hyperbolic discounting functions proposed in the literature, we only give this general characterisation that makes explicit the property that the near future in clock-time is associated with relatively larger time distance evaluation than the far future.

For both exponential and hyperbolic discounting, note that there is a degree of freedom involved in the choice of the function *Disc* that gives the discounting factors. In the above results, it has been assumed that *Disc* behaves so as to preserve the correspondence between clock-time and time distance. Yet, for an infinite horizon, it is possible that this is not the case: it is easy to see that converse cases could hold in which there is no proportionality (or concavity) in the correspondence, yet *Disc* is chosen in a way that gives an exponential (or hyperbolic) time discounting function.

4.4.6 Discounted Value

We will now discuss how the evaluation of temporal features established by the above framework can be combined with a goodness evaluation to obtain a representation of discounted value. This will make transparent that additional assumptions are needed to endorse the time discounting of *value*. That is to say, while the representation of discounting developed so far allows for a completely separate and general evaluation of time distance in temporally extended prospects, additional assumptions are needed in order to establish time discounted value.

In a first step, we have to assume that it is possible to obtain a goodness evaluation of the events in Q, such as by a variant of standard expected utility theory. Note that in order to apply the discounting weights, a (weak) separability assumption is required: it has to be the case that a value can be assigned to any

$q \in Q$, and that values of prospects can be obtained by combining the evaluations of several events. With these assumptions in place, discounted value can be established by joint consideration of the two evaluations, i.e. the time distance evaluation and the goodness evaluation.

Secondly, recall that the time discounting function as introduced above gives weights that are completely independent of the goodness evaluation. Crucially, the descriptive and normative status of discounted goodness will depend on the time distance interpretation given. Indeed, a set of assumptions is needed to establish that discounted value is descriptively and normatively meaningful. That is, the degree to which one accepts the normative and descriptive attractiveness of a particular interpretation of time distance determines the degree of the descriptive and normative appeal of discounted utility. More generally, the initial separation of time and value makes transparent the requirements for establishing discounted value in both a descriptive and normative value.

4.4.7 Summary

This section has provided a general foundational framework for time discounting. We asked what assumptions are needed to establish well-founded time discounting functions, and showed that they need to satisfy both a representation and a correspondence requirement. More specifically, in the measurement-theoretic framework for time discounting provided in this section, a discounting factor can be determined by a representation of time distance between events. It has been shown that if such a representation corresponds to an externally given time-index, a general discounting function according to Definition 1 can be recovered.

In a general sense, the framework renders transparent the formal and conceptual assumptions required by theories of time discounting. That is, the general framework developed here has a number of applications in foundational work regarding time discounting. Formally, the framework can be employed to assess and render transparent formal assumptions that specific accounts of time discounting make. Conceptually, it can be related to a number of interpretations of time distance, including time preference, preference change, delay as well as risk and uncertainty. From an empirical point of view, it can be asked whether existing accounts of descriptive time discounting approaches satisfy the measurement conditions needed to specify their functional form. Concerning a possible justification of time discounting, the framework lends itself to a neutral comparison

of the normative appeal of different substantial interpretations of time distance.

Naturally, not all of those possible applications can be discussed in the context of this thesis. We will proceed by revisiting the four problems of time discounting as well as time preference theories in Section 4.5, before discussing how the general framework for time discounting given here can be motivated by the concept of connectedness in the multiple-self in Section 4.6.

4.5 Time Discounting Theories Reconsidered

This section applies the general foundations of time discounting to (i) the four problems of time discounting introduced earlier, and (ii) time preference theories of time discounting. Before discussing these two problems, we reconsider the main insights from the general foundations of time discounting developed in Section 4.4.

Firstly, the general framework (initially) separated the time and the value dimensions. That is, we have discussed foundations of time discounting by asking what kind of evaluation of temporal features of prospects they can capture. The separate analysis has a twofold motivation: firstly, if time discounting factors are to be used to weight goodness evaluations, the former should live up to requirements of measurement and representation that are comparable to those on which goodness evaluations are founded. Secondly, it was argued that time discounting representations that entangle time and value make it more difficult to discuss the four problems of time discounting. We will argue below that separating time and value in the general framework enables us to ask well-posed questions about time discounting concerning the four problems.

Secondly, the account distinguished between an evaluation of temporal features of events (the time distance) and an externally given time-index (the clock-time). This exposed a curious dichotomy inherent in time discounting functions. On the one hand, in order for time discounting factors to be well-founded, the numerical values have to be derived or justified axiomatically. In the general framework, this was achieved by the representation of time distance. On the other hand, time discounting functions have to give weights that are associated with an external-time index in a quite regular way. In the general framework, this has been considered by studying different kinds of correspondence between clock-time and time distance. That is to say, in the terminology of the general

framework, the twofold concern of time discounting is addressed by the characterisation of 'representation' and 'correspondence', respectively. We will argue below that the joint requirement of correspondence and representation puts severe constraints on time discounting theories.

In a general sense, the discussion suggests that the framework developed here makes explicit the formal and conceptual requirements for time discounting theories, casting doubts on whether current proposals of time discounting theories can meet those requirements.

4.5.1 Four Problems of Time Discounting

Here we reconsider the four problems of time discounting raised earlier from the perspective of the general foundations of time discounting. Recall that the four problems of time discounting concern the descriptive and normative discussion of two questions: firstly, what is the correct time discounting function, and secondly, what is the correct conceptual motivation for time discounting?

Functional Form

Here we discuss the question of the correct functional form of discounting, in both a descriptive and in a normative sense. The general framework allows us to discuss specific function forms of time discounting by formulating conditions on the correspondence between clock-time and time distance. This was exemplified by Theorems 14 and 15 which state exponential and hyperbolic time discounting in the general framework, respectively. Note that conditions formulated on the correspondence between clock-time and time distance imply a specific behaviour of the representation of time distance. That is, the deserved functional form of time discounting can be conveniently expressed in terms of correspondence, which fixes requirements for the representation.

In Theorem 14, it is shown that exponential time discounting requires that time distance is proportional to clock-time (i.e. a linear transformation of clock-time yields time distance). That is, assuming a discrete set Q for ease of exposition, the time distance between any two events that are subsequent in clock-time must be of equal difference: for example, the time distance between events in 2 and 3 years will be the same as the time distance between events in 22 and 23 years, and so on. This very strong condition on the representation of time distance is required for proportionality of correspondence between clock-time and

time distance. Descriptively, this condition is probably best understood as an approximation, while it is hard to see a direct normative appeal for it. Yet, arguments about those questions will depend on what kind of interpretation of time distance is considered.

In Theorem 15, non-exponential time discounting is characterised by time distance that is concave in clock-time (such that a concave transformation of clock-time yields time distance). The condition is relevant for hyperbolic discounting as it can be used to describe the phenomenon that time distance of events that are close to the present in clock-time is perceived as relatively larger than the time distance of events that are in the far future in clock-time. Again assuming a discrete set Q for ease of exposition, this means that the time distance between, say, events in 2 and 3 years will be the larger than the time distance between events in 22 and 23 years. That is, the concavity condition allows one to evaluate the time distance of events that have intervals of equal length in clock-time differently. This condition is more permissive than linearity, which implies a uniformity in time distance evaluation. As such, it can be seen as a more plausible condition in a descriptive sense, while its normative appeal will again depend on the interpretation of time distance that is adopted.

Now recall the two features of the general framework mentioned above. Firstly, consider the joint requirement of correspondence and representation. The above discussion shows that the functional form of discounting can be expressed in terms of correspondence, yet that these in turn imply conditions on the representation. This makes explicit the requirement that specific functional forms of time discounting can only be justified when the representation of time distance is well-behaved enough to satisfy the correspondence conditions. As will be discussed later in more detail, the joint requirement of correspondence and representation severely constrains the kinds of interpretations that can be used to conceptually motivate time discounting.

A comparison of the above joint requirement to the kinds of requirements implied by representations common in expected utility theory further illustrates the latter fact: in expected utility theory, all what is required is a representation of preference by a function with interval-scale properties. In addition to such a requirement, time discounting representations also require ratio-scale properties of time distance (for normalisation to the present) as well as correspondence between the time distance representation and clock-time (i.e. increasing correspondence

for a general time discounting function, as well as more specific conditions, such as linearity or concavity for exponential or hyperbolic discounting). As a loose analogy, consider how expected utility is sometimes linked to expected monetary value, such as in the law of diminishing marginal utility. Here, monetary value serves as an externally given index to which utility is linked in a regular way. Yet, such considerations are certainly outside the confines of (contemporary) decision theories and the concept of utility as preference satisfaction. That is to say, the general framework developed here exposes the fact that the requirements for time discounting are much more complex than those for related concepts, such as those for expected utility.

Secondly, the separation between time and value in the general framework shows how the normative appeal of exponential discounting is weaker than usually presumed. As briefly mentioned when initially considering the four problems of time discounting, there is an argument for exponential discounting due to its preservation of the utility function over time. That is, exponentially discounted utility will stay stable over time. By contrast, hyperbolically discounted utility can yield inconsistency over time: weighting equal clock-time intervals in the near future different than those in the far future implies that such weightings can change over time. That is, as time passes, the weightings assigned to time points can also change, as the far future will become the near future and can hence be evaluated hyperbolically again. This is usually taken to be an argument for exponential time discounting, where such problems do not arise. Yet, the general framework shows that this argument is based almost exclusively on considerations of value. It does not show that there is something inherently wrong with, for instance, a concave correspondence between clock-time and time distance. All it suggests is that such an evaluation can undermine goodness evaluations. The latter can still be taken as an argument for the normative appeal of exponential discounting. Yet, the argument is an additional consideration motivated by the preservation of goodness, and not automatically motivated by the representation of time distance. This weakens the argument considerably, as it does not show that there is anything inherently problematic about non-linear correspondence between clock-time and time distance.

Conceptual Motivation

Here we discuss the underlying conceptual motivation for time discounting. In the general framework, the representation of time distance was introduced for which a conceptual motivation is required in order to motivate discounting.

Consider the question of motivating time discounting in a descriptive sense. In this context, the general framework makes precise the regularity assumptions that are needed for time discounting: possible time discounting motivations such as time impatience, delay perception, or preference change can only be used to interpret the framework if they behave in such a regular fashion that they are representable. Indeed, for a well-founded time discounting function, the axioms contained in an ordinal distance structure have to be motivated. Moreover, as discussed in the previous section, specific conditions on the correspondence between clock-time and time distance imply further regularity constraints on the representation of time distance.

The general framework also permits us to formulate well-posed questions about the elicitation of agents' attitudes to time distance. Indeed, after fixing a particular interpretation of time distance (for instance, one associated to time impatience, or changes in propositional attitudes), axioms on ordinal distance could also be tested in elicitation exercises where agents compare time distances according to such an interpretation between pairs of events. Given that the ordinal distance measurement procedure can be understood as separate from goodness evaluations, this will allow us to determine whether there are interpretations of time distance that allow us to delineate agent's attitudes towards time distance from those to goodness in such a practical sense. It is beyond the scope of this work to discuss such more practical problems of eliciting agents' attitudes to time distance in sufficient detail. Still, the virtue of the general measurement procedure developed here with regards to the elicitation of agents' attitudes to time distance lies in posing the elicitation task in a sparse and well-defined way: if it is possible to elicit an agent's attitudes to time distance according to ordinal distance measurement, then a time discounting can be based on 'revealed ordinal time distance'. This is an improvement over representations that entangle time and value, as it exposes the requirement of time discounting to delineate attitudes to time distance from other attitudes of agents.

Normative discussion about time discounting usually rests on the idea that it should not undermine goodness evaluation, as discussed in the previous sec-

tion. Here, the separation of time and value in the general framework exposes the requirement of a separate normative justification of time discounting. That is to say, a conceptual motivation that underpins discounting that fulfils this requirement would answer the question why agents should engage in discounting the future.

4.5.2 Time Preferences

Here we show how the general framework is capable of reformulating the time preference approach. As reviewed earlier, time preference theories of discounting obtain discounted value by considering a preference relation $\succsim_{TP}$ on an outcome-time structure, which can be represented by a discounted utility function.

In order to reformulate this idea, recall that the general framework conceives of discounted value of prospects as the combination of two *separate* evaluations, a goodness evaluation and a time distance evaluation D of events that form prospects. The goodness evaluation can be given in a standard decision-theoretic framework, where the goodness evaluation maximises a preference relation on events, and a time distance evaluation can be given by a variant of the general framework of time discounting which combines a representation of time distance $\succcurlyeq$ with a correspondence requirement.

Time preferences, to be reformulated in this account, have therefore to be 'separated' into two components, one concerning value and one concerning time. The value (or goodness) dimension is reasonably straightforward, as it can be given by a standard expected utility account. The time dimension requires us to reformulate two conditions in outcome-time structures, namely time impatience and stationarity. (The latter is only needed when endorsing exponential time discounting, which we will assume here.)

Concerning time impatience, the relation $\succcurlyeq$ in ordinal distance structures can be interpreted as ordering pairs of events according to the impatience they invoke, where $rp \succcurlyeq qp$ iff a higher degree of impatience is associated with r than with q, as no time impatience is associated with p. Isolating time impatience in this fashion highlights the fact that the conceptual assumptions in time preference theories have been, by and large, implicit in the frameworks. As discussed in the initial review of conceptual motivations for time discounting, there are a number of different ideas of what time impatience consists in and how it is 'produced', such as by a desire for immediate gratification, or by cognitive deficiencies ('weakness

of imagination'). These conceptual nuances will matter a great deal in deciding whether time impatience can motivate a representation of time distance in an ordinal distance structure.

Concerning stationarity, this condition can be mirrored by the linearity requirement that underlies the formulation of exponential time discounting in Theorem 14. Recall that stationarity requires time preferences to be invariant under equal distances of clock-time (that is, the ordering of two outcomes at clock-time t and $t+\mu$ remains the same when considering the two outcomes at t' and $t'+\mu$, respectively). This assertion is captured in linearity of correspondence between clock-time and time distance. Note how this reformulation, analogous to time impatience, is now completely separate from the goodness evaluation.

Such a reformulation of time preference accounts exposes the fact that the exponential shape of time discounting functions often endorsed by those theories is ambiguous in its conceptual underpinning (as it primarily stems from the idea to preserve the goodness evaluation). Indeed, it is unclear how time impatience implies a regularity of time distance so as to motivate linearity of correspondence between clock-time and time distance. In a general sense, the framework developed here allows us to reconsider the assumptions made about time and goodness evaluations in a separate way to expose what kinds of formal and conceptual requirements are encapsulated in time discounting theories.

4.6 Time Discounting in the Multiple-Self

This section presents multiple-self interpretations of the general discounting framework introduced in the previous sections. Firstly, Parfit's proposal that time discounting can be motivated by diminishing intrapersonal connectedness between temporal selves is used to interpret the time distance. On his account, events $q \in Q$ are interpreted as richer propositions that are relative to temporal selves. The greater the similarity of preferences between temporal selves, the closer the discounting factor is to the unit weight. We show that 'Parfit's representation theorem' for time discounting is helpful in addressing objections against his view put forward by Williams (1970) and Elster (1986). Secondly, we suggest that Parfit's discounting can motivate a novel account of (exponential) time discounting for preference change.

4.6.1 Parfit's Discounting for Intrapersonal Connectedness

As reviewed in Chapter 3, Parfit (1984) proposes to view personal identity over time reductively, and as a matter of psychological connectedness. Furthermore, psychological connectedness is viewed as a matter of degree, such that different temporal instances of a person can be connected to each other to a varying degree of strength. This very idea of imperfect intrapersonal connectedness can, according to Parfit, also motivate time discounting. More specifically, Parfit (1984, 313) maintains:

> 'My concern for my future may correspond to the degree of connectedness between me now and myself in the future ... since connectedness is nearly always weaker over long periods, I can rationally care less about my further future.'

The quote suggests that we can view persons as consisting of collections of temporal selves that are connected to each other in varying degree. This, in turn, motivates time discounting. The above claim I will call Parfit discounting, using it to interpret the ordinal distance structure of the general time discounting framework developed in the previous sections.

Recall an ordinal distance structure $\langle Q \times Q, \succcurlyeq \rangle$. For Parfit discounting, the set Q can be interpreted as a collection of events which have a richer description such that they include a reference to a temporal self. For example, the description of the event of having desert includes a reference to a specific self, such as 'self S_i is eating desert'. As before, the relation $\succcurlyeq$ orders *pairs* in this set. The dimension of the distance comparison is the psychological distance between the temporal selves associated with the events. That is to say, the relation $\succcurlyeq$ orders pairs of events according to the similarity of the psychological traits of the temporal selves that are associated with those events. That is, for all events $q, r, s, t \in Q$, if $qr \succ st$ then the psychological differences between the selves that are associated with q and r are greater than those between the selves that are associated with s and t. This suggests that it is possible to obtain a numerical representation of psychological connectedness between selves. Furthermore, we can interpret the general representation framework for time discounting given in Section 4.4 with Parfit's intrapersonal connectedness. Accordingly, if each self in the multiple-self corresponds to a clock time-point, we can motivate time discounting factors according to intrapersonal connectedness. Hence, Parfit discounting can be used to motivate a

time discounting function that satisfies Definition 1. What is needed in order for the time discounting representation to hold conceptually, is that ordinal distances between selves can be compared according to their psychological connectedness. It is hard to see why comparisons of psychological similarity between pairs of temporal selves are a stronger requirement than distance comparisons according to, for instance, time impatience. Indeed, recalling the review of time impatience theories, it can be argued that Parfit's account of psychological connectedness is a more comprehensive source of conceptual motivation that the notion of time impatience in time preference theories. This suggests that Parfit discounting is at least as viable a candidate theory for time discounting as the ones discussed so far.

Parfit discounting for psychological connectedness is meant to be a normative concept. However, Frederick (1999) present an empirical study in which subjects' intrapersonal connectedness has been tried to establish via questionnaires, asking participants to make connectedness comparisons between their future selves. While he is by and large pessimistic about the possibilities that this method provides robust empirical evidence for time discounting factors, his approach suggests that the measurement-theoretic framework can indeed be interpreted by intrapersonal connectedness, as it even lends itself to elicitation exercises.

Problems with Parfit Discounting

Parfit discounting has been criticised for a variety of reasons. Here, we consider four arguments against his proposal.

The first line of criticism focuses almost exclusively on the fact that Parfit's psychological connectedness is not the correct characterisation of personal identity over time, as briefly reviewed in Chapter 3. The worry is that Parfit connectedness is a non-starter concerning personal identity, and therefore of no use in answering other questions. Without claiming to fully rebuke such a highly sceptic stance, consider the discussion of multiple-self models in Chapter 2, and their possible interpretation with conceptions of intrapersonal connectedness, as reviewed in Chapter 3. In this discussion, we suggested that connectedness in the multiple-self is introduced as a modelling device to investigate the relation between time and decision-making, rather than as a complete account of personal identity. In that sense, it is still possible to maintain that Parfit connectedness can yield an interpretation of a specific aspect of personal identity, without covering all

aspects of it. This, in turn, makes it possible to use his account to motivate time discounting.

The second line of criticism suggests that Parfit connectedness cannot influence a concern for future selves. In other words, Parfit connectedness does not imply Parfit discounting, as many other kinds of connectedness will play a role in determining concern for future selves. The most prominent proponent of such a view is Williams (1970) who maintains that imperfect intrapersonal connectedness does not disburden us from concern for our future selves. Indeed, Williams (1970) asks us to imagine a situation in which there is low psychological connectedness, yet there is still connectedness in terms of the body, memories and sympathy. He argues that in such a situation, we should still be concerned with a future self thus connected to our present self. It seems that Williams does not want to reject Parfit discounting, but rather suggest that his account of connectedness is unduly narrow, and that, for instance, bodily connectedness, or some other type of non-reductive connectedness maintains perfect overall intrapersonal connectedness, and with it maintains concern for future selves. That is to say, it is not only a matter of (dis-)similarity of preference between temporal selves that can motivate time discounting: for instance, consider someone who knows he will have, by and large, different preferences when approaching retirement age. That is, there is low psychological connectedness between his self now and the self at retirement age. Were we to discount goods for that future self according to psychological connectedness, then a very low discounting factor would apply. Yet, the present self might still feel empathy for his future self, and might still care about whether the future self will receive goods. This, in turn, suggests that richer accounts, modifying connectedness to interpret the ordering relation $\succcurlyeq$ with a non-reductive account as well, would be needed to fully answer this criticism.

The third line of criticism appeals to the notion of rational agency and maintains that Parfit discounting is in conflict with it. Rational agency, this argument goes, overrides any imperfections in concern for the future that may arise out of diminishing psychological connectedness. Prominently, the rejection of time discounting by Sidgwick (1907) and Rawls (1971) that was mentioned in Section 4.2.4 is already associated with a statement of this position, namely by Rawls (1971, 259), who said that 'rationality requires an impartial concern for all parts of our life.' Closely related to this assertion, Elster (1986) claims that agency uni-

fies temporal selves despite imperfect connectedness. Other proponents of this view are, for instance, Korsgaard (1989), as already briefly discussed in Chapter 3. On this view, even if Parfit's connectedness account is accepted, the discounting that is motivated by it is in conflict with a theory of goodness that suggests that there has to be equal concern for selves.

Yet, this critique is misplaced if we consider Parfit discounting in the general representational framework. In the latter, due to the separation of time and value, psychological connectedness functions solely as an enrichment, and does not replace any of the features of a rational agent. Rather, it offers a modelling device to consider the influence of time distance on goodness evaluations, and it is unclear why it is problematic that time distance can have an effect on goodness evaluations. Consider the discussion of temporal distance in standard decision theories in Chapter 2. We have shown that standard decision-theoretic accounts stay, by and large, silent on the issue of temporal distance and do indeed allow for future events to be valued less than current ones. What they do not offer, are tools in which temporal distance can be evaluated separately which is what the representational framework for time discounting, combined with a psychological connectedness interpretation offers here.

Furthermore, such an enrichment is minimal when considering psychological connectedness. Assume that a theory of goodness is consequentialist. In a first step, the goodness of a prospect can be evaluated with, for instance, expected utility theory. Now, Parfit discounting does provide a theory of time discounting which uses variations in the very basis of evaluating goodness as its starting point: it proposes time discounting on the basis that psychological features will be different in the future. In other words, Parfit discounting measures the present credibility of the goodness evaluation (if the latter is based on preferences). That means that Parfit discounting does not diminish the concern for future selves, it just expresses the degree to which future selves share the present evaluations. This is arguably an important concern for the present evaluations: Parfit discounting simply expresses to what degree one can expect present evaluations to be diachronically stable.

A fourth line of criticism, raised by Parfit (1984) himself and by Elster (1986), is that discounting for diminishing psychological connectedness has unfortunate implications: it takes future wants as datum, rather than something that can be shaped, and it does not respect the fact that there are some things about

the future that we value deeply, such as the fulfilment of plans. The following statement, hence, would be ruled out by and be vulnerable to this objection. Parfit (1971, 9):

> 'We care less about our further future because we know that less of what we are now – less, say, of our present hopes or plans, loves or ideals – will survive into the further future. ... [if] what matters holds to a lesser degree, it cannot be irrational to care less.

In this quote, Parfit takes psychological connectedness to be a very rich concept, which includes plans and ideals. However, note that such a rich interpretation of psychological connectedness cannot be used to motivate an empirical structure in a time distance representation, as it would not be compatible to the regularity properties in ordinal distance structures, as given in Definition 3. More specifically, it is not clear how it would be possible that individual's hopes, plans, or ideals that are associated with events can be ordered according to the relation $\succcurlyeq$. The kind of psychological connectedness that *can* motivate a time distance representation is one of psychological similarity between temporal selves, as determined by sameness of taste, or preference. Such interpretations can also fulfil the correspondence requirement in a representational framework. That is to say, if we wish to motivate a well-founded time discounting function, it cannot be as richly motivated as implied by the above quote. Hence, this line of criticism seems to rest on a misunderstanding of the formal confines of time discounting.

It seems, then, that the viability of Parfit's representation theorem rests on a narrow reading of his idea of psychological connectedness as taste or preference similarity between temporal selves. Yet, this does not imply that plans, hopes, or ideals have to be disregarded. It only suggests that they cannot be adequately captured by time discounting functions and should be considered separately. Indeed, plans, hopes, and ideals can still inform a separate goodness evaluation. Consider an individual who chooses her career path and could either become a banker or pursue graduate studies in physics. Now, this person might have the ambition to become a scientist. Clearly, such an ambition will be reflected in the goodness evaluation of the two available prospects. Since the prospects are extended through time, the decision-maker might also consider her psychological connectedness and discount the expected goodness of both prospects accordingly.

What is required for the latter, however, is that independence between connectedness and the action in question holds. That is, if there is an action available

to the decision-maker that is designed to change her psychological connectedness, then discounting comes in conflict with the evaluation of such an action. Such an independence requirements mirrors conditions for probabilistic independence of acts and states in Bayesian decision theories. While it somewhat limits the applicability of psychological connectedness to motivate time discounting, it presents a far smaller limitation than the initial objection against discounting for psychological connectedness suggested.

4.6.2 Exponential Discounting for Taste Change

To emphasise the capabilities of the general framework, this section briefly considers a novel motivation for exponential time discounting that is closely related to Parfit discounting. Indeed, upon Parfit discounting satisfying linear correspondence we can motivate exponential discounting for psychological connectedness. Such a regular taste change might be plausible if we take it as an approximation of the psychological changes an individual undergoes. Indeed, if an individual deliberates about a time horizon in which little of her circumstances change, a quite regular taste change might be defended. This is not to say that it does not present a heavy assumption – yet, it is hard to see why it is a stronger assumption than, say, a regular behaviour of time impatience. Furthermore, making the assumption of linear correspondence yields a conceptually interesting derivation of exponential discounting.

Note that exponential discounting is usually derived from time preference accounts which assume stable tastes and preferences. The reason for the latter assumptions is that time preferences are defined on outcome-time structures, and a constant discounting factor only holds if tastes are stable over time. Such requirements are not needed in the general framework, where time and value are separated. Formally, it is possible to derive time discounting by using any concept that can motivate the time distance representation and correspondence requirements.

Hence, we may assume Parfit connectedness as the motivation for time discounting, and if it satisfies linear correspondence, we have an account of exponential time discounting that is motivated by the changing psychological characteristics of temporal selves within the decision-maker. Moreover, as discussed when introducing the concept of psychological connectedness in Chapters 2 and 3, psychological connectedness can be understood as a coarse-grained character-

isation of changes in taste.[8] After deriving exponential time discounting in such a fashion, we can combine it with a standard goodness evaluation which yields exponentially discounted utility, while allowing for changing tastes.

Nevertheless, we may ask whether combining the two accounts in this way gives rise to conflicts. Firstly, we might say that goodness evaluation and time evaluation really are separate procedures and it is possible to make opposing assumptions. Secondly, the interpretation of preference change in the two evaluations is rather different. When evaluating goodness, a single preference change can cause a preference reversal which undermines an agent's utility function. However, when evaluating time according to psychological connectedness, only coarse-grained changes in tastes are considered. That is, an agent can have an unconditional belief about how likely his tastes are going to change in the future, for instance, by introspection about the past development of his personality. Indeed, if there would be more structure, the regularity assumptions in the ordinal distance structure and linear correspondence could not be satisfied. This suggests that there is no conflict between evaluating goodness of prospects according to present (and stable) preferences on the one hand and then weight those evaluations with an evaluation of how likely such preferences are to change over time.

More generally, consider that other connectedness interpretations according to Chapter 3 are possible. This yields a rich set of possibilities to motivate time discounting according to psychological and empathy connectedness (as well as other ones due to reductive and non-reductive memory connectedness).

Note that motivating an ordinal distance representation in the above fashion can also be interpreted as providing measurement-theoretic foundations for connectedness functions in multiple-self models. That is to say, instead of assuming that connectedness is a matter of degree, the ordinal distance representation yields a connectedness function that can be transformed into giving well-founded connectedness weights. In the remainder of the thesis, when employing discounting weights that are motivated by connectedness, we shall assume that such a foundation can be given in principle, such that the connectedness weights are well-founded. This will be especially useful in Chapter 6.

[8]Note that the latter does not refer to changes in preference about specific, identifiable propositions as in preference reversals or dynamic inconsistency (those kinds of preference changes will be analysed separately in Chapter 6).

4.7 Conclusions

This chapter has provided a representational framework for time discounting. In this framework, a discounting factor can be based on a representation of time distance that corresponds to an externally given time-index. Formally, this provides a structure to assess the temporal element of prospects in intertemporal decisions. Conceptually, it can be related to a number of interpretations of time differences, including time impatience, risk and uncertainty, intrapersonal connectedness, and preference change.

The analysis in this chapter has shown that the concept of time discounting has a specific role in intertemporal decisions: it can provide weights that quantify qualitative properties associated with the passage of time, to provide an account of how temporal distance can be evaluated. For this to hold, such qualitative properties have to be measurable by some procedure (ideally, some variant of the general framework outlined in Section 4.4.2). Note how measurability by such a procedure severely constraints the conceptual motivations for time discounting functions. Many complex issues, such as how decision-makers form plans for the future, and ambiguities about their execution, cannot possibly be fully (if at all) captured by time discounting. The kind of evaluation of intertemporality which can be performed by time discounting is one of a coarse-grained evaluation of features of temporal distance. That is, time discounting functions *can* be motivated by the idea that the passage of time is in general associated with such concepts as impatience, with additional risk or uncertainty, or a general change in an individual's personal identity. However, time discounting functions do not lend themselves to characterize many complexities that are related to the evaluation of intertemporal prospects. The chapter thus clarifies the concept of time discounting by rendering explicit the fact that a construction of time discounting factors requires us to endorse strong regularity conditions.

4.8 Appendix: Proofs

Proof of Theorem 5

To prove Theorem 5, we first consider that ordinal distance structures as given in Definition 3 are closely related to so-called 'absolute-difference structures' in Krantz *et al.* (1971, 170ff.), henceforth called KLST-type structures. For the latter, Krantz *et al.* (1971, 173) prove a representation similar to Theorem 5. In KLST-type structures, the following conditions are different from the ones listed in Definition 3:

2. Symmetry. If $q \neq r$, then $qr \sim rq \succ qq \sim rr$.

3. Well-Behavedness.

   (i) If $r \neq s$, $qs \succcurlyeq qr, rs$ and $rt \succcurlyeq rs, st$, then $qt \succcurlyeq qs, rt$.

   (ii) If $qs \succcurlyeq qr, rs$ and $qt \succcurlyeq qs, st$, then $qt \succcurlyeq rt$.

In order to include elements that are distinct yet identical under the dimension of comparison, the above conditions have been replaced by a weak symmetry and a reformulated well-behavedness condition, respectively, in the ordinal distance structure given in Definition 3. To show how ordinal distance structures relate to KLST-structures, we define a binary relation $\equiv$ on Q by $q \equiv r$ iff $qr \sim qq$, for all $q, r \in Q$. Interpretationally, such $q \equiv r$ means that q and r are identical with regards to the dimension of comparison (such as sweetness, or time distance, etc.).

Lemma 16. *If $\langle Q \times Q, \succcurlyeq \rangle$ is an ordinal distance structure, then $\equiv$ is an equivalence relation on Q. Moreover, this equivalence relation is congruent with $\succcurlyeq$, i.e. for all $q, r, s, t, q', r', s', t' \in Q$, if $q \equiv q'$, $r \equiv r'$, $s \equiv s'$, $t \equiv t'$, then $qr \succcurlyeq st$ iff $q'r' \succcurlyeq s't'$.*

Proof. Reflexivity: $q \equiv q$ because $qq \sim qq$ (since $\sim$ is reflexive). Symmetry: If $q \equiv r$ then $qr \sim qq$, and so (by weak symmetry) $rq \sim rr$; so that $r \equiv q$. Transitivity: Suppose $q \equiv r$ and $r \equiv s$. Then $qr \sim qq$ and $rs \sim rr$. By $qr \sim qq$, we have $qs \sim rs$ (using weak symmetry), which by $rs \sim rr$ and $rr \sim ss$ implies $qs \sim ss$, i.e. $q \equiv s$. Congruence: Suppose $q \equiv q'$, $r \equiv r'$, $s \equiv s'$, $t \equiv t'$. We prove $qr \succcurlyeq st$ iff $q'r' \succcurlyeq s't'$. Applying weak symmetry four times, we have $qr \sim q'r \sim q'r'$ and $st \sim s't \sim s't'$ So $qr \succcurlyeq st$ iff $q'r' \succcurlyeq s't'$. □

By Lemma 16, for every ordinal distance structure $\langle Q \times Q, \succcurlyeq \rangle$ we can define a new structure $\langle Q^* \times Q^*, \succcurlyeq^* \rangle$ called "$\langle Q \times Q, \succcurlyeq \rangle$ modulo equivalence $\equiv$":

- Q^* is the set $Q/\equiv$ of all equivalence classes of Q with regards to $\equiv$,
- $\succcurlyeq^*$ is the binary relation on $Q^* \times Q^*$ defined as follows: for all $q^*, r^*, s^*, t^* \in Q^*$, $q^*r^* \succcurlyeq s^*t^*$ iff $qr \succcurlyeq st$ for some (hence, by congruence, any) elements q, r, s, t from the equivalence classes q^*, r^*, s^*, t^*, respectively.

Lemma 17. *If $\langle Q \times Q, \succcurlyeq \rangle$ is an ordinal distance structure, then $\langle Q^* \times Q^*, \succcurlyeq^* \rangle$ is a KLST-type structure (Krantz* et al.*, 1971, 170).*

Proof. The properties of $\langle Q \times Q, \succcurlyeq \rangle$ imply the KLST-type properties, by drawing on congruence (which allows us to replace any relation $q^*r^* \succcurlyeq s^*t^*$ by $qr \succcurlyeq st$ for any representants q, r, s, t of q^*, r^*, s^*, t^*, respectively) and by noting that:

- in the KLST Condition 2, we can replace $q^* \neq r^*$ by $qr \succ qq$ (which, by adding the weak symmetry condition 2.(ii) in Definition 3, implies $qr \sim rq \succ qq \sim rr$),
- in the KLST Condition 3.(i) we can replace $r^* \neq s^*$ by $rs \succ rr$ (which reduces the condition to the Condition 3 in Definition 3),
- KLST Condition 3.(ii) can be omitted as it is implied by the new Weak Symmetry condition. To show this, let (a) $qs \succcurlyeq qr, rs$ and (b) $qt \succcurlyeq qs, st$. It has to be shown that $qt \succcurlyeq rt$. By completeness of $\succcurlyeq$, $rs \succ rr$ or $rr \succcurlyeq rs$.

  Firstly, suppose $rs \succ rr$. If $rt \succcurlyeq rs, st$, the result follows directly from (b) and WBi. Assume (c) $rs \succ rt$ or (d) $st \succ rt$. If (c), then $qs \succcurlyeq rt$ by (a) and transitivity of $\succcurlyeq$. By (b) and transitivity therefore $qt \succcurlyeq rt$. Assume (d). Then, by transitivity and (b), $qt \succcurlyeq rt$.

  Secondly, suppose $rr \succcurlyeq rs$. Since also $rs \succcurlyeq rr$ (by Weak Symmetry (i)), it follows that $rs \sim rr$. Hence, by Weak Symmetry (ii), $st \sim rt$. This, together with (b) and transitivity implies that $qt \succcurlyeq rt$.

□

We can now prove Theorem 5.

Proof. Suppose $\langle Q \times Q, \succcurlyeq \rangle$ is an ordinal distance structure. Then, by Lemma 17 above, $\langle Q^* \times Q^*, \succcurlyeq^* \rangle$ is a KLST-type structure. By Krantz *et al.* (1971, 173), there is a function $\varphi^* : Q^* \to \mathbb{R}$ such that φ^* represents $\succcurlyeq^*$. Define $\varphi(q) := \varphi^*(q^*)$, where q^* is the equivalence class of q. For all $q, r, s, t, \in Q$, we have $qr \succcurlyeq st$ iff $q^*r^* \succcurlyeq s^*t^*$ iff $|\varphi^*(q^*) - \varphi^*(r^*)| \geq |\varphi^*(s^*) - \varphi^*(t^*)|$ iff $|\varphi(q) - \varphi(r)| \geq |\varphi(s) - \varphi(t)|$. □

Proof of Corollary 6

Proof. Immediate from the properties of the interval representation in Theorem 5. □

Proof of Proposition 7

Proof. Consider any $q, r \in Q$. We have

$$\begin{aligned} rp \succcurlyeq qp &\Leftrightarrow |\varphi(r) - \varphi(p)| \geq |\varphi(q) - \varphi(p)| \text{ (as } \varphi \text{ represents } \succcurlyeq) \\ &\Leftrightarrow |\varphi(r)| \geq |\varphi(q)| \text{ (by } \varphi(p) = 0) \\ &\Leftrightarrow \varphi(r) \geq \varphi(q) \\ &\Leftrightarrow Disc \circ \varphi(r) \leq Disc \circ \varphi(q) \text{ (as } Disc \text{ is decreasing).} \end{aligned}$$

□

Proof of Proposition 9

Proof. Suppose c and c' are two correspondences. To show that they coincide, consider any clock-time $t \in T$. By assumption, there is an event $q \in Q$ such that $\tau(q) = t$. Since c and c' are correspondences, $c \circ \tau(q) = c' \circ \tau(q)$, i.e., $c(t) = c'(t)$. □

Proof of Proposition 10

Proof. 1. First suppose that (*) for all $q, r \in Q$, $\tau(q) = \tau(r) \Leftrightarrow \varphi(q) = \varphi(r)$.

Define a function $c : T \to \mathbb{R}$ as follows. For all $t \in T$ let $c(t) = \varphi(q)$, where q is a (by assumption existing) event in Q with clock time $\tau(q) = t$. By (*), this definition does not depend on the choice of q. By definition, $\varphi = c \circ \tau$. So, c is a correspondence.

2. Conversely, suppose there exists a correspondence c. We have to show (*). Consider any events $q, r \in Q$ such that $\tau(q) = \tau(r)$. Applying c on both sides

it follows that $c \circ \tau(q) = c \circ \tau(r)$, and hence by $\varphi = c \circ \tau$ that $\varphi(q) = \varphi(r)$, as desired. □

Proof of Proposition 11

Proof. 1. First suppose that

(*) for all events $q, r \in Q$, $\tau(q) > \tau(r) \Leftrightarrow \varphi(q) > \varphi(r)$.

To show that c is increasing, consider any clock-times $t, s \in T$ such that $t > s$. Write them as $t = \tau(q)$ and $s = \tau(r)$ for some events $q, r \in Q$. We have $c(t) = c \circ \tau(q) = \varphi(q)$ and $c(s) = c \circ \tau(r) = \varphi(r)$. By $\tau(q) > \tau(r)$ and (*), we have $\varphi(q) > \varphi(r)$. Hence, by $\varphi = c \circ \tau$, we have $c \circ \tau(q) > c \circ \tau(r)$, i.e., $c(t) > c(s)$. This proves that c is increasing.

2. Conversely, suppose that c is increasing. We have to show (*). Consider any events $q, r \in Q$ such that $\tau(q) > \tau(r)$. Since c is increasing, it follows that $c \circ \tau(q) > c \circ \tau(r)$, in other words, that $\varphi(q) > \varphi(r)$, as desired. □

Proof of Theorem 12

Proof. Conditions (ii) and (iii) are equivalent by Proposition 11 above. So it suffices to show that (i) and (ii) are equivalent.

1. First, suppose (ii). We have to show that D has range $(0, 1]$, is decreasing and satisfies $D(0) = 1$. Since $Disc$ has range $(0, 1]$, so does D. Since $Disc$ is decreasing and c increasing, the composition D is decreasing. Finally,

$$D(0) = Disc \circ c(0) = Disc(0) = 1.$$

So, D is a time discounting function.

2. Now suppose (i). Applying the $Disc^{-1}$ on both sides of $D = Disc \circ c$, we obtain $Disc^{-1} \circ D = c$. So c is the composition of two decreasing functions, and hence is increasing, proving (ii). □

Proof of Theorem 14

Proof. Let $Disc$ preserve linearity by $Disc(i) = \delta^i$ for all $i \in I$, and let c be linear, say $c(t) = k$ for some $k \in \mathbb{R}$. If $T = \{0\}$ the result holds trivially. So let $T \neq \{0\}$. It follows that $k > 0$, since otherwise c would not be increasing. For all $t \in T$, $D(t) = Disc(c(t)) = \delta^{kt} = (\delta^k)^t$. Note that $\delta^k < 0$ by $k > 0$. So D is exponential with discount factor δ^k. □

Proof of Theorem 15

Proof. Let *Disc* preserve concavity by $Disc(i) = \delta^i$ for all $i \in I$, let c be concave in clock time, and $\#T > 2$. For a contradiction, suppose D is exponential, say $D(t) = \gamma^t$ for some $0 < \gamma < 1$. For all $t \in T$, we have $D(t) = Disc(c(t))$ and hence $\delta^t = \gamma^{c(t)}$. Taking logarithms on both sides, we get $t \log \delta = c(t) \log \gamma$, and hence $c(t) = kt$ for $k := \frac{\log \delta}{\log \gamma}$. So c is linear. This implies that c is not concave. To see why, note that by $\#T > 2$ we can find distinct $t, t' \in T$ and an $a \in (0,1)$ such that $at + (1-a)t' \in T$. We have

$$\begin{aligned} c(at + (1-a)t') &= k(at + (1-a)t') \\ &= akt + (1-a)kt' \\ &= ac(t) + (1-a)c(t'), \end{aligned}$$

whereas concavity would require that $c(at + (1-a)t') > ac(a) + (1-a)c(t')$. This contradiction completes the proof. □

Chapter 5

Backward Induction

Summary. This chapter analyses the problem of interaction over time; in particular, the sequential structure of dynamic games with perfect information. A three-stage account is proposed, that specifies set-up, reasoning and play stages of dynamic games. Accordingly, we define a player as a set of agents corresponding to these three stages. Moreover, the notion of agent connectedness is introduced which measures the extent to which agents' choices are sequentially stable. A type-based epistemic model is augmented with agent connectedness and used to provide sufficient conditions for backward induction. Moreover, an existence result is obtained ensuring that these conditions are indeed possible. Our epistemic foundation for backward induction makes explicit that the epistemic independence assumption involved in backward induction reasoning is stronger than usually presumed. Furthermore, in the three stage-account, players can explicitly be understood as multiple-selves, which permits to interpret low agent connectedness as stemming from imperfect connectedness between selves.[1]

5.1 Introduction

Strategic interaction over time is modelled by dynamic games. The standard extensive form models dynamic games as trees, but does not further explicate the sequential dimension. The structure of the game, the players, their reasoning and strategies are implicitly assumed to remain stable throughout the whole game. In particular, the reasoning is supposed to occur before the game and to apply

[1]This chapter is based on a joint paper (Bach and Heilmann, 2009) with Christian W. Bach (University of Maastricht, Netherlands) to which both authors contributed equally.

to the entire duration of the game. However, local deviations from strategies are relevant for the dynamics of sequential interaction. More specifically, agents may depart from the strategy of their respective player, thus contradicting the idea that agents act according to instructions. Here, we perceive of a player as a set of agents and introduce the notion of *agent connectedness* to capture the extent of sequential stability of players. In our account, high agent connectedness characterises an agent's compliance with a player and low agent connectedness an agent's deviation. Precisely such properties of agents are central to backward induction, since players need to be able to entertain deviating moves by opponents' agents in hypothetical reasoning. Indeed, here we provide sufficient conditions for backward induction in terms of agent connectedness, as well as an existence result ensuring that our conditions are indeed possible.

In a general sense, we amend the representation of a dynamic game by three sequential stages. In the set-up stage, the game structure and the players' utilities are determined. Then, in the reasoning stage, the players deliberate about the game, their opponents and choose their strategies. Finally, in the play stage, the players' agents act at their respective decision nodes. Relative to these three sequential stages, a player is defined as a set of agents, namely the set-up agent, the reasoning agent and the game agents. We also amend the notion of strategy such that its use in the stages can be discussed separately, introducing the notion of initial strategy in the reasoning stage and actual strategy in the play stage. This three-stage account enables us to make explicit the sequential stability assumptions inherent in dynamic games. Also, the framework can be used to locally weaken such assumptions.

The reasoning of players in games is usually described by epistemic models. Here, we extend a type-based epistemic model of dynamic games with an initial strategy function by means of which the connectedness of each agent to his respective player can be expressed. In particular, the notion of connectedness between a player's reasoning agent and his game agents is formally introduced to capture the assumption of sequential strategic stability, i.e. compliance with the initial strategy. According to this definition, an agent is either high-connected if he acts in line with the initial strategy or low-connected otherwise. Hence, beliefs about the connectedness of opponents' agents enters the belief space as an additional epistemic feature. Applying this framework, sufficient conditions for backward induction are obtained by explaining surprise information with low-

connectedness of the deviating agent. Rather than revising the belief in an opponent's rationality, a supposedly irrational move of one of his agents at a preceding decision node is accommodated by belief revision on the high-connectedness of that agent. This, in turn, separates that supposedly irrational agent from the remaining agents of the respective opponent.

Various substantial interpretations of this framework become available. Interpreting sequence temporally, the three stages in a dynamic game reflect a player as existing over time: initially, a player assigns utilities to possible outcomes, subsequently chooses a strategy and at later points in time, he actually plays. Indeed, players existing over time can be interpreted as multiple-selves and their agents as selves whose personal identity changes over the course of the game. This, in turn, makes available theories that describe how personal identity over time can change. In particular, the behavioural notion of agent connectedness in a player can then be explained by intrapersonal connectedness. We would like to stress that *agent connectedness* as introduced here is not directly integrated into a multiple-self model as introduced in Part I of this thesis. Rather, agent connectedness is a behavioural notion that refers to a game agent's compliance (high agent connectedness) or deviation (low agent connectedness) from the initial strategy. We will discuss in Section 5.5 how this behaviour can be explained by characterisations of underlying connectedness in a multiple-self model of personal identity over time.[2]

Furthermore, the interpretation of our framework unveils strong assumptions implicit in the principle of epistemic independence which underlies any foundational argument for backward induction. Indeed, an observed surprise must never induce a belief revision on any intrapersonal connectedness of game agents at any later points in the game. To illustrate our framework, consider the dynamic game with perfect information given by the extensive form given in Figure 5.1.

Such games are commonly solved by backward induction as follows. At *Alice*'s second decision node, her unique optimal choice is f. Given this choice of *Alice*, *Bob*'s unique optimal action at his decision node is d. Given the unique optimal choices of *Alice* at her second decision node and *Bob* at his decision node, *Alice* picks a at her first decision node. The backward induction strategy profile (af, d) thus obtains. Note that *Bob* has to entertain the possibility of *Alice* having

[2]In the remainder of this chapter, the term connectedness will be used to refer to the behavioural notion of agent connectedness. Any reference to (underlying) connectedness as used in the multiple-self model will be made explicit.

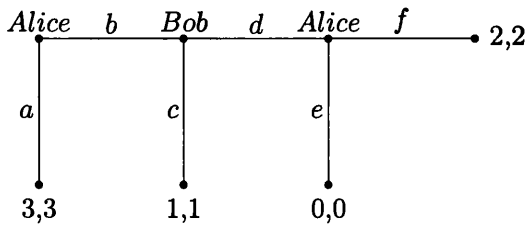


Figure 5.1: A Dynamic Game with Perfect Information

deviated from her backward inductive strategy when determining the choice for his decision node. Even though *Alice* could not have complied with her backward inductive strategy, *Bob* is assumed to think that *Alice* will nevertheless act in accordance with backward induction later on and hence to play the backward inductive move himself at his decision node. Accounting for the surprise that Alice has played b while still maintaining that she will play d is vital in making backward induction reasoning work. Usually, an assumption of epistemic independence is used to exclude any influence of such deviating behaviour on expectations about *Alice*'s future behaviour. In our account, surprise information is explained with low agent connectedness of the agent governing *Alice*'s first decision node. Low agent connectedness can be interpreted by understanding player *Alice* as a multiple-self. Accordingly, the deviating behaviour of *Alice*'s first agent can be explained as exhibiting low underlying connectedness between *Alice's* selves. For instance, such low intrapersonal connectedness can occur due to a breakdown of psychological features, memory or empathy between the self at *Alice*'s first decision node and her other selves. Such a more detailed description of dynamics allows us to more fully understand strategic interaction over time.

This chapter proceeds as follows. Section 5.2 introduces a three-stage account of dynamic games, and defines a player as a set of agents. Section 5.3 describes the players' reasoning by extending a type-based epistemic model of dynamic games with agent connectedness. Section 5.4 gives sufficient conditions for backward induction in terms of agent connectedness and provides an existence result ensuring the possibility of these conditions. Section 5.5 discusses some interpretative issues of our framework, with particular emphasis on a multiple-self interpretation of a player. Finally, Section 5.6 offers some concluding remarks.

5.2 Modelling Dynamic Games

Since Kuhn (1953) dynamic strategic interaction has commonly been modelled by the so-called extensive form which represents a game as a tree.

Definition 1 (Extensive form structure with perfect information). *An* extensive form structure with perfect information *is a tuple* $\Gamma = (X, Z, E, x_0, I, m, (u_i)_{i \in I})$, *where*

- X *is a finite set of non-terminal nodes, specifying decision nodes,*
- Z *is a finite set of terminal nodes, specifying the different situations in which the game may end,*
- E *is a finite set of directed edges* $(x, y) \in X \times (X \cup Z)$, *specifying the choices for the players, where* (x, y) *moves the game from* x *to* y,
- x_0 *is the unique root of the tree and called initial node,*
- I *is a finite set of players, where* $\mid I \mid > 1$,
- $m : X \backslash Z \rightarrow I$ *is the move function assigning to every non-terminal node the choosing player, where* X_i *denotes the set of all* $x \in X$ *such that* $m(x) = i$,
- $u_i : Z \rightarrow \mathbb{R}$ *is player* i*'s utility function assigning to every terminal node* $z \in Z$ *a utility* $u_i(z)$.

The extensive form can be interpreted as a set-up procedure for modelling a dynamic game. Note that we restrict attention to dynamic games with perfect information, i.e. games in which all players, whenever they have to choose, know exactly the choices made by their opponents until then. First, the structure of the game has to be specified, i.e. all of its possible situations, outcomes at final situations, and rules are formalised by the sets X, Z and E, respectively. Then, a particular set of players determines the decision-makers in the game and the corresponding contingent situations where they act is given by the move function m. In a final step, each player has to consider all possible outcomes of the game and assign cardinal utilities to them in line with his preferences. To make explicit this procedural character inherent in the extensive form, we call the course of fixing the model the *set-up stage* of a dynamic game. Once the game is fixed, the players can reason about it for decision-making purposes and thereafter the game is actually played.

A further basic ingredient when modelling dynamic games is the notion of strategy, which is considered the object of choice for the players. A strategy specifies an action for each contingency that might possibly arise for the respective player.

Definition 2 (Strategy). *Let Γ be an extensive form structure with perfect information, $i \in I$ some player and $A_i(x_i) \subseteq E$ the set of edges departing from $x_i \in X_i$. A* strategy *for i is a function $s_i : X_i \to \bigcup_{x_i \in X_i} A_i(x_i)$ such that $s_i(x_i) \in A_i(x_i)$ for all $x_i \in X_i$.*

According to the standard view, a strategy specifies an action for each contingency that might possibly arise for the respective player and hence can be interpreted as his disposition to act at each of his decision nodes. We call such a choice plan *actual strategy* since it refers to the contingent actions of the players when actually playing the game. However, also before the game is played, players determine strategies based on their hypothetical reasoning. Such objects resulting from the players' reasoning and being fixed before play are called *initial strategies* and are formally defined in the next section. Note that actual strategies can differ from initial strategies.

Indeed, after the set-up stage, a player reasons about his opponents as well as the fixed game, and decides on a complete contingent choice plan for the game as a result of this reasoning. We call this process the *reasoning stage* of a dynamic game and the player's ensuing hypothetical choice plan is his initial strategy. Note that although coming after the set-up stage, the reasoning stage is prior to the actual play of the game. The introduction of the reasoning stage thus explicitly separates hypothetical plans from actual choices.

After the set-up and reasoning stages the game is actually played and all contingent situations that may possibly arise in the game are represented in the extensive form by a player's set of decision nodes. We assume that each such node is governed by an agent of the player and call the actual playing phase of the dynamic game the *play stage*. With the game structure and initial strategy being fixed by the prior two stages, the play stage determines the strategy profile that is actually played as well as the corresponding outcome and utilities for the players. Hence, our account distinguishes between three stages of a dynamic game: the set-up, the reasoning and the play stages.

Further, note how our three-stage view on dynamic games makes use of the notion of player. Accordingly, two distinguishable tasks are performed by a player

before the play stage: utilities have to be assigned to outcomes in the set-up stage, followed by the choice of an initial strategy in the reasoning stage. During the play stage each of the decision nodes specifies a distinguishable task to be handled by one agent, respectively. In order to be able to discern the acting entities of the different stages, we understand the player as consisting of a *set-up agent*, a *reasoning agent* and *game agents*. Formally, a player is defined as the set of his agents.

Definition 3 (Player as set of agents). *Let Γ be an extensive form structure with perfect information. A player $i \in I$ in Γ is defined as a set of agents $i = \{\alpha_s^i, \alpha_r^i, \alpha_1^i, \alpha_2^i, ..., \alpha_m^i\}$, where $\mid X_i \mid = m \in \mathbb{N}$, and α_s^i is called* set-up agent, *α_r^i is called* reasoning agent, *and all other agents α_j^i are called* game agents, *each corresponding to a unique decision node $x_i \in X_i$.*

Accordingly, a player as a set of agents makes formally explicit the different tasks to be performed by a player in a dynamic game, related to the three different stages. The conception of a player as set of agents can naturally be linked to the notion of selves in a multiple-self. Indeed, the idea of understanding agents of players as distinct selves of multiple-selves has been used in the context of dynamic games with imperfect information by, for instance, Piccione and Rubinstein (1997). Also, the idea that a player consists of different acting selves appears in the agent normal form in Selten (1975). In Definition 3, the agents of a player are modelled according to their specific tasks in the game. However, from an interpretative point of view, a more detailed description of agents in a player can be considered.

Note that our account of dynamic games makes transparent their sequential structure. Yet, a stability assumption lurks in the standard extensive form model. Despite the sequential character of dynamic games, no changes in the game's ingredients, utility assignments or choice prescriptions by the initial strategy is admitted during the dynamic interaction. Indeed, any object fixed in the two pre-play stages, once determined, remains rigid until the end of the game. In particular, the deliberation of the reasoning agent of a player is supposed to apply to all game agents, who are all required to adhere to the initial strategy. Hence, any dynamics concerning the game structure as well as concerning the players are excluded by the standard extensive form model of dynamic games. While it may seem plausible to keep the game structure fixed given an underlying dynamic game to be modelled, the suspension of any dynamics concerning the

players represents a rather strong assumption within the standard model. The introduction of three stages relevant to a dynamic game allows us to explicitly endorse or weaken the stability assumption with respect to deviation from pre-play reasoning.

The idea of understanding a player as a set of agents is now illustrated with the extensive form depicted in Figure 5.1. In addition to the game agents at their respective decision nodes, both players *Alice* and *Bob* have two further agents that determine their utilities and strategies before play, corresponding to the set-up and reasoning stage, respectively. The two players can thus be formalised as sets $Alice = \{Alice_s, Alice_r, Alice_1, Alice_3\}$ and $Bob = \{Bob_s, Bob_r, Bob_2\}$. Actual choice of a player is then described by a strategy, each component of which is determined by the respective game agent in charge. For example, the actual strategy profile (be, d) signifies that $Alice_1$ chooses b at her first decision node, then Bob_2 picks d at his decision node and $Alice_3$ selects e at her second decision node. However, the initial strategies of the reasoning agents could be different. For instance, $Alice_r$ might have chosen bf prior to play. Note that in this example, a common index is used for both players to identify the position of their agents in the game tree and to reflect its sequential structure. More complicated game trees such as ones with parallel nodes governed by different agents of one player can then still be given some sequential order, relative to the structure of the game tree. Further, game agents assigned to decision nodes that are excluded by actual play can be interpreted as inactive game agents. Also, it is possible to conceive of a player as having inactive agents at opponent decision nodes, and to hence interpret the player as a decision-maker over time with inactive agents at points in the game where no game agent acts for him. For instance, the set representing *Alice* would then be amended with the inactive agent $Alice_2$, and the set representing *Bob* would be amended with the inactive agents Bob_1 and Bob_3, where the inactive agents correspond to decision nodes which are assigned to opponent game agents, respectively.

Our three-stage account of dynamic games proposed in this section makes explicit the sequential character of dynamic games and the stability assumptions already implicit in the standard extensive form model. A player is conceived of as a set of agents relative to the three sequential stages, making explicit that different agents of a player act in distinct sequential situations before and during play. The next section proposes an epistemic model for the reasoning stage of

dynamic games and formalises the notion of initial strategy.

5.3 Extending Type-based Interactive Epistemology

Interactive epistemology, also called epistemic game theory when applied to games, constitutes a quite recent field of inquiry into the foundations of game theory. Interactive epistemology provides an abstract framework to formalise epistemic notions such as belief and knowledge in settings involving several decision-makers. This field of research was initiated by Aumann (1976) and first adopted in the context of games by Aumann (1987) as well as Tan and Werlang (1988). Developments of the discipline are reviewed by, for instance, Brandenburger (1992), Battigalli and Bonnano (1999), Board (2002), Brandenburger (2007), and Perea (2011, forthcoming). The fundamental problem addressed is the description of the players' choices in a given game relative to epistemic assumptions. For instance, a typical starting point for epistemic game-theoretic analysis is the notion of common knowledge of rationality.[3] Implications for play in given classes of games are then deduced. More generally, existing game-theoretic solution concepts are characterised in terms of epistemic assumptions as well as novel solution concepts are proposed by studying the consequences of refined or new epistemic hypotheses.

Epistemic game theory can be regarded as complementing classical game theory. While classical game theory is based on two basic primitives – game form and choice – epistemic game theory adds an epistemic framework as a third basic component, on the basis of which knowledge and beliefs can be modelled in games. Moreover, the epistemic program in game theory is somewhat opposed to the classical refinement program. The latter approach takes the Nash (1951) notion of equilibrium as the starting point and attempts to propose various refinements, that cut down the multiplicity of equilibria with the ultimate objective of obtaining a unique prediction for play in an arbitrary game. In contrast to seeking a general characterisation of rationality in terms of equilibrium refinement, the epistemic programme takes beliefs, knowledge and rationality as starting points for its analysis and aims to unveil the implications of epistemic assumptions for play in games. A wide variety of epistemic hypotheses involving different refinements of rationality as well as different epistemic operators can be considered and their subsequent game-theoretic consequences be analysed. As Brandenbur-

[3] An event E is common knowledge among a set of agents G if E holds, every $i \in G$ knows E, every $i \in G$ knows that every $i \in G$ knows E, ... (ad infinitum).

ger remarks, a key property of the epistemic programme is that there does not exist one right set of assumptions to make about a game (Brandenburger, 2007, 490). In particular, the concept of Nash equilibrium loses its predominant role, since it merely qualifies as one particular outcome of a game – obtainable only under specific epistemic conditions – among a set of possibilities. For a more detailed comparison of classical game theory and the epistemic programme, see de Bruin (2009).

More generally, epistemic game theory builds on the basic intuition that a player has to reason about the other players. Before choosing his strategy, he must form a belief about what his opponents will do. However, in order to so, he also needs to form a belief about what the others believe that their opponents will do. Similarly, any higher-order beliefs about his opponents are relevant to the player's choice. In order to formally represent players' reasoning about each other, an epistemic model is added to the classical analysis of a game.

Here, we follow the type-based approach to epistemic game theory, according to which different epistemic states are encapsulated in the notion of type. Note that the notion of type was originally introduced by Harsanyi (1967) in the specific context of incomplete information but can actually be generalised to any interactive uncertainty. A recent survey of type-based interactive epistemology is provided by Siniscalchi (2008). In a type-based epistemic framework for games – as illustrated in Example 4 – a set of types is assigned to every player, where each player's type induces a belief on the opponents' choices and types. Thus, any higher-order belief can be derived from a given type.

Example 4. *Let* $\Gamma = ((S_i)_{i\in I}, (u_i)_{i\in I})$ *be a game in normal form, where* $I = \{Alice, Bob\}$ *is a set of players,* $S_{Alice} = \{x, y, z\}$, $S_{Bob} = \{0, 1\}$ *are their strategy sets, and* u_i *are utility functions for* $i \in \{Alice, Bob\}$. *The tuple* $\mathcal{M}^\Gamma = ((T_i)_{i\in I}, (b_i)_{i\in I})$ *is an epistemic model of* Γ, *where* T_i *is a finite set of types for* $i \in \{Alice, Bob\}$ *and every type* $t_i \in T_i$ *induces a probability function* $b_i(t_i) \in \Delta(S_i \times T_i)$ *on the opponents' choice-type combinations. Suppose the following types and believes:*

- *Sets of types:*

  $T_{Alice} = \{t^1_{Alice}, t^2_{Alice}, t^3_{Alice}\}$ *and*

  $T_{Bob} = \{t^1_{Bob}, t^2_{Bob}, t^3_{Bob}, t^4_{Bob}\}$

- *Beliefs for Alice:*

$b_{Alice}(t^1_{Alice}) = (1, t^2_{Bob})$

$b_{Alice}(t^2_{Alice}) = (0, t^1_{Bob})$

$b_{Alice}(t^3_{Alice}) = .3 \cdot (0, t^2_{Bob}) + .7 \cdot (1, t^3_{Bob})$

- *Beliefs for Bob:*

$b_{Bob}(t^1_{Bob}) = (y, t^2_{Alice})$

$b_{Bob}(t^2_{Bob}) = (x, t^1_{Alice})$

$b_{Bob}(t^3_{Bob}) = (z, t^3_{Alice})$

$b_{Bob}(t^4_{Bob}) = .5 \cdot (x, t^1_{Alice}) + .5 \cdot (z, t^2_{Alice})$

Note that any higher-order belief can now be derived for the different types of a player from the epistemic model. For instance, Alice's type t^1_{Alice} *believes with probability* 1 *that Bob chooses* 1. *Also,* t^1_{Alice} *believes with probability* 1 *that Bob believes with probability* 1 *that Alice picks* x. *Moreover,* t^1_{Alice} *believes with probability* 1 *that Bob believes with probability* 1 *that Alice believes with probability* 1 *that Bob selects* 1. *Analogously, any higher-order belief can be derived for* t^1_{Alice}.

To give another example, Bob's type t^2_{Bob} *believes with probability* 1 *that Alice chooses* x. *Also,* t^2_{Bob} *believes with probability* 1 *that Alice believes with probability* 1 *that Bob picks* 1. *Moreover,* t^2_{Bob} *believes with probability* 1 *that Alice believes with probability* 1 *that Bob believes with probability* 1 *that Alice selects* x. *Analogously, any higher-order belief can be derived for* t^2_{Bob}.

To give an example with partial beliefs, Alice's type t^3_{Alice} *believes with probability* .3 *that Bob chooses* 0; *and with probability* .7 *that Bob chooses* 1. *Also,* t^3_{Alice} *believes with probability* .3 *that Bob believes with probability* 1 *that Alice picks* x; *and* t^3_{Alice} *believes with probability* .7 *that Bob believes with probability* 1 *that Alice picks* z. *Moreover,* t^3_{Alice} *believes with probability* .3 *that Bob believes with probability* 1 *that Alice believes with probability* 1 *that Bob selects* 1; *and* t^3_{Alice} *believes with probability* .7 *that Bob believes with probability* 1 *that Alice believes with probability* .3 *that Bob selects* 0 *and that Alice believes with probability* .7 *that Bob selects* 1. *Analogously, any higher-order belief can be derived for* t^3_{Alice}.

In a similar fashion, for every type of every player, any higher-order belief can be obtained from the epistemic model.

Here, we extend the standard type-based epistemic model with the new notion of initial strategy. Before our epistemic model can be defined, one further notion

is needed. Letting S_j denote the set of all strategies of player j, a strategy $s_j \in S_j$ is said to avoid a given decision node $x \in X$, if there exists some decision node $x^* \in X$ on the unique path from the initial node x_0 to x, for which s_j assigns an off-path action. The set $S_j(x) \subseteq S_j$ then denotes all strategies of player j that do not avoid node x. An extended epistemic model for dynamic games can now be defined as follows.

Definition 5 (Extended epistemic model). *Let Γ be a finite extensive form structure with perfect information. An* extended epistemic model *of Γ is a tuple $\mathcal{M}^\Gamma = (T_i, \beta_i, \iota_i)_{i \in I}$, where*

- *T_i is a finite set of types for player i,*
- *$\beta_i : T_i \times (X_i \cup x_o) \to \Delta(\times_{j \in I \setminus \{i\}}(S_j \times T_j))$ assigns to every type $t_i \in T_i$ and decision node $x_i \in X_i$ a probability distribution on the set of opponents' strategy-type pairs, where $\beta_i(t_i, x_i) \in \Delta(\times_{j \in i \setminus \{i\}}(S_j(x) \times T_j))$ for all $x \in X_i \cup \{x_0\}$,*
- *$\iota_i : T_i \to S_i$ assigns to every type $t_i \in T_i$ an initial strategy.*

In the context of our three-stage account of dynamic games, the extended epistemic model concerns the reasoning stage. Thus, the deliberation of a player's reasoning agent is formalised by the extended epistemic model. In particular, the reasoning agent is disposed with conditional beliefs of any order at each of his decision nodes as well as the initial node, via the probability function β_i. Crucially, it is a distinguished feature of our epistemic model that a type does not only hold conditional beliefs about the opponents' *actual* strategies, but also about the opponents' *initial* strategies. Note that the conditional beliefs of the reasoning agent refer to hypothetical epistemic states of the respective game agents. Hence, while types and their induced conditional belief hierarchies model the deliberation process of the reasoning agent, the novel ingredient of initial strategy is interpreted as the outcome of the player's reasoning. Further, note that the conditional beliefs of a player i at a given node $x \in X \cup \{x_0\}$ only assign positive probability to opponents' strategy choices that do not avoid x. This seems reasonable since otherwise a player would exhibit contradictory beliefs: although knowing to be at decision node x, he believes that at least one opponent has chosen a strategy avoiding x and thus excluding it to be reached.

The initial strategy is fixed in the reasoning stage before the play stage, in which the game is actually played. Choices by the game agents might differ from

the ones prescribed by the reasoning agent's initial strategy. Since a player is conceived of as a set of agents by Definition 3, such a behavioural deviation from the initial strategy raises the problem of connectedness between a player's game agents and his reasoning agent. On the basis of the initial strategy function, we now formally introduce connectedness into our extended epistemic model for dynamic games.

Definition 6 (Agent connectedness). *Let $\mathcal{M}^\Gamma$ be an extended epistemic model of an extensive form structure Γ with perfect information. Further, let $\iota_i^{t_i}(x_i)$ denote the action that the initial strategy of type $t_i \in T_i$ designates for game agent $\alpha^i_{x_i}$ at $x_i \in X_i$, let s_i^α denote the strategy of i that is actually played and let $s_i^\alpha(x_i)$ denote the actual choice of game agent $\alpha^i_{x_i}$ at x_i. The* agent connectedness $c_i(\alpha^i_{x_i}, s_i^\alpha \mid t_i)$ *of game agent $\alpha^i_{x_i}$ is defined as*

$$c_i(\alpha^i_{x_i}, s_i^\alpha \mid t_i) = \begin{cases} \textit{high} & \textit{if } \iota_i^{t_i}(x_i) = s_i^\alpha(x_i), \\ \textit{low} & \textit{otherwise.} \end{cases}$$

In Definition 6, the actual strategy played refers to the actual choices of the respective player's game agents at the decision nodes they govern. Initial and actual strategy are then compared. A game agent is said to be *high-connected* if he acts in compliance with the initial strategy and *low-connected* otherwise. Connectedness hence both separates and relates sequential parts of the player at contingent points of the game. Note that the connectedness function expresses a behavioural notion as its values are determined by the actual choices of the game agents, relative to the initial strategy of the reasoning agent. In this context, the reasoning agent can be seen as the central representative of the player. This is plausible as the reasoning agent initially chooses a complete strategy that is intended to apply throughout the game, whereas the game agents only act locally. Also, note that stability of the initial strategy and hence equivalence to the actual strategy is implicitly assumed in the standard extensive form model. In the sequel, we therefore refer to reasoning agent and player interchangeably.

Moreover, the notion of agent connectedness in Definition 6 can be explained by underlying connectedness of psychological features, memory and empathy as captures by multiple-self models of personal identity over time. Also, rather than focusing on the relation between the reasoning agent and the game agents of a given player, connectedness between any pair of agents can be considered. Such interpretations are addressed in Section 5.5.

In type-based epistemic models, the objects of beliefs are events. Intuitively, an event states a property concerning the model's uncertainty space. Within the context of games, examples of events are "*Alice* plays strategy bf", "*Bob* is rational" and "*Bob* believes at the initial node that *Alice*'s agents are high-connected". Formally, events are simply sets of types. More precisely, a set $E \subseteq \bigcup_{i \in I} T_i$ of types is called event. The belief of some player i at some node $x \in X_i \cup \{x_0\}$ in some event E can then be modelled by projecting $\beta_i(t_i, x_i)$ on T_{-i}, denoted as $\beta_i(t_i, x_i \mid T_{-i})$. Similarly, player i's belief at some node $x \in X_i \cup \{x_0\}$ on the type of player $j \in I \setminus \{i\}$ can be obtained by projecting $\beta_i(t_i, x_i)$ on T_j, denoted as $\beta_i(t_i, x_i \mid T_j)$. Moreover, player i's belief at some node $x \in X_i \cup \{x_0\}$ on player j's strategy-type pair can be extracted by projecting $\beta_i(t_i, x_i)$ on $S_j \times T_j$, denoted as $\beta_i(t_i, x_i \mid S_j \times T_j)$. Note that beliefs are events, too and that indeed any higher-order belief can be represented in a type-based epistemic model. Given some event, a player's type specifies conditional belief hierarchies at each of his decision nodes. Epistemic states are thus local and concern the respective node-governing agent of the player. Yet they are hypothetical in the sense of belonging to the reasoning agent when deliberating before play about what his game agents would know were their respective nodes be reached.

For the purpose of formalising rationality in our framework, let $u_i(\iota_i(t_i), \beta_i(t_i) \mid x_i)$ denote player i's expected utility starting at node x_i of playing the relevant part of strategy $\iota_i(t_i)$ given his belief at x_i about the opponents' strategies.

Definition 7 (Rationality). *Let $\mathcal{M}^\Gamma$ be an epistemic model of an extensive form structure Γ with perfect information and $i \in I$ some player. A type $t_i \in T_i$ is* rational *if $t_i \in R_i = \{t_i \in T_i : u_i(\iota_i(t_i), \beta_i(t_i) \mid x_i) \geq u_i(s_i, \beta_i(t_i) \mid x_i)$ for all $s_i \in S_i$ and for all $x_i \in X_i\}$.*

Accordingly, a type of a player is rational if his initial strategy maximises his expected utility at every decision node in the game. Rationality is hence understood as a notion relative to the result of a player's reasoning, since it is precisely the outcome of his reasoning that reflects his attitude towards the interactive situation he is involved in. Note that our notion of rationality is weaker than the standard one, since the latter requires actual choice to be optimal throughout the tree while the former only concerns initial choice. A player can thus be rational in our sense while still actually acting irrationally in the standard sense. As an illustration of this observation, consider the game given in Figure 5.1. Suppose *Alice* believes that *Bob* chooses d and her reasoning agent $Alice_r$ hence picks the

rational initial strategy af. Nevertheless, $Alice_1$ can still choose b at her decision node, hence acting irrationally in the standard sense.

As has already been pointed out above, within our extended epistemic framework a player can be perceived as the reasoning agent and his object of choice, the initial strategy, can be perceived as the result of the reasoning process, for instance, as his intention or plan of action for the game. A player's game agents then actually choose actions at the respective decision nodes, either in line with the initial strategy of the reasoning agent, or differently. Specific patterns of relationship between a player or his reasoning agent and his game agents can be formalised. We call a player $i \in I$ *high-connected* if all of his game agents are highly connected, i.e. $c_i(\alpha^i_{x_i}, s_i^\alpha \mid t_i) = high$ for all $\alpha^i_{x_i} \in i$. In other words, the game agents of a high-connected player actually choose in complete accordance with his proposed initial strategy. However, it is possible that only some game agents are highly connected, while others are not. For instance, only game agents succeeding some particular node might be high-connected. Crucially, belief in different patterns of high-connected game agents can be defined in our model. For instance, the following condition requires a player to believe in the high-connectedness of an opponent at all future nodes.[4]

Definition 8 (Future-high-connectedness). *Let $\mathcal{M}^\Gamma$ be an epistemic model of an extensive form structure Γ with perfect information, $i \in I$ be some player, and $x \in X_i \cup \{x_0\}$ some node. A type $t_i \in T_i$* believes in j's future-high-connectedness *at node x if $t_i \in BH_{i,j}(x) = \{t_i \in T_i : supp(\beta_i(t_i, x \mid S_j \times T_j)) \subseteq \{(s_j, t_j) \in S_j \times T_j : s_j(x_j) = \iota_j^{t_j}(x_j)$ for all $x_j \in X_j$ succeeding $x\}\}$.*

Accordingly, a player's belief on what an opponent is actually playing is related to the belief about his initial strategy. More generally, our model is also capable of distinguishing between actual and initial strategy in the reasoning of players about their respective opponents.

The preceding definitions introduce connectedness of game agents to their player, conceived of as the reasoning agent, into our extended epistemic framework, which in turn can be used to understand reasoning in strategic interaction over time. In a first such step, these notions are used in the next section to shed light on backward induction reasoning in dynamic games with perfect informa-

[4]The following definition uses the mathematical notion of support abbrevieated as *supp*. The support of a probability measure on some space X is the set containing all elements $x \in X$ that receive positive probability.

tion.

5.4 Sufficient Conditions for Backward Induction

Backward induction constitutes the standard reasoning method in dynamic games with perfect information: at each decision node, optimal behaviour is determined by assuming the optimality of choices at all succeeding nodes. Before formally defining backward induction we restrict attention to generic games with perfect information.

Definition 9 (Genericity). *An extensive form structure Γ with perfect information is called* generic *if for every player $i \in I$, for every decision node $x_i \in X_i$, for every two actions $a_i, a_i' \in A_i(x_i)$, every two terminal nodes $z \in Z$ such that z follows a_i and z' follows a_i', it holds that $u_i(z) \neq u_i(z')$.*

Accordingly, any two different choices at a given decision node will always lead to two distinct utilities for the respective player. It is common to assume genericity when searching for epistemic characterisations of backward induction. Since genericity implies uniqueness of the backward inductive strategy profile, no ambiguity arises in determining the actions in line with backward induction at each node in the tree. This restriction is not severe, since the aim is to unveil the epistemic states portraying the way of thinking characteristic of backward inductive reasoning. Genericity avoids the introduction of somewhat arbitrary criteria for ties that would divert from the essential properties of the players' reasoning required for backward induction to obtain.

Further note that backward induction can only be defined for finite games, as possible end points of the game are required for the backward inductive process to begin. Finiteness is already implicit in our definition of the extensive form.

In order to facilitate the formal expression of backward induction, the decision nodes are classified according to their maximal distance from an end point i.e. a terminal node of the game, independent from any closer terminal nodes.

Definition 10 (Decision nodes). *Let Γ be an extensive form structure with perfect information and $x \in \bigcup_{i \in I} X_i$ some decision node. Decision node x is called* ultimate *if x is only immediately succeeded by terminal nodes; decision node x is called* pre-ultimate *if x is only immediately succeeded by ultimate decision nodes or by ultimate decision nodes and terminal nodes; decision node x is called* pre-pre-ultimate *if x is only immediately succeeded by pre-ultimate decision nodes*

or by pre-ultimate decision nodes and ultimate decision nodes or by pre-ultimate decision nodes and terminal nodes or by pre-ultimate decision nodes and ultimate decision nodes and terminal nodes; etc. Decision node x is called initial node *if* $x = x_0$.

It is now possible to define backward induction for generic finite dynamic games of perfect information as follows.

Definition 11 (Backward induction strategy). *Let Γ be a generic extensive form structure with perfect information, $i \in I$ some player and $x_i \in X_i$ some decision node of i. The unique backward inductive choice $b_i(x_i) \in A_i(x_i)$ at x_i is determined as follows: if x_i is an ultimate node, then $b_i(x_i)$ is the unique action that maximises i's utility at x_i, and if x_i is pre-ultimate node, then $b_i(x_i)$ is the unique action at x_i that maximises i's utility at x_i given backward inductive actions at all decision nodes succeeding x_i, etc. Player i's unique* backward inductive strategy *$b_i \in S_i$ assigns to each of i's decision node $x_i \in X_i$ the respective unique backward inductive action $b_i(x_i) \in A_i(x_i)$.*

Naturally, epistemic game theory searches for epistemic requirements that induce the players to choose their backward inductive strategies. Indeed, various different sufficient conditions for backward induction have been proposed in the literature, which are reviewed, unified and compared by Perea (2007). Furthermore, note that the emphasis lies on what requirements are needed for a player to actually *choose* his backward inductive strategy and hence to make transparent the complete reasoning underlying backward induction. The genuinely different question of what epistemic conditions are needed to get the backward inductive *outcome* is addressed in, for instance, Battigalli and Siniscalchi (2002) and Brandenburger *et al.* (2008). Here, we give an epistemic characterisation of the backward inductive strategy profile in terms of connectedness.

Some more epistemic concepts need to be introduced before formal conditions for backward induction can be stated.

Definition 12 (Structural belief in rationality). *Let $\mathcal{M}^\Gamma$ be an epistemic model of an extensive form structure Γ with perfect information, and $i \in I$ some player. A type $t_i \in T_i$* structurally believes in his opponents' rationality *if $t_i \in SBR_i = \{t_i \in T_i : supp(\beta_i(t_i, x \mid T_j)) \subseteq R_j$, for all $x \in X_i \cup \{x_0\}$, for all $j \in I \setminus \{i\}\}$.*

Accordingly, at the beginning of the game as well as at any of his decision nodes,

a player believes that all of his opponents are rational i.e. choose a rational initial strategy.

Iterating structural belief in rationality gives the nested epistemic notion of common structural belief in rationality.

Definition 13 (Common structural belief in rationality). *Let $\mathcal{M}^\Gamma$ be an epistemic model of an extensive form structure Γ with perfect information and $i \in I$ some player. A type $t_i \in T_i$ expresses* common structural belief in rationality *if $t_i \in CSBR_i = \{t_i \in T_i : t_i \in SBR_i^k$ for all $k \geq 1\}$, where $SBR_i^1 = SBR_i$, and $SBR_i^{k+1} = \{t_i \in T_i : supp(\beta_i(t_i, x \mid T_j)) \subseteq SBR_j^k$, for all $x \in X_i \cup \{x_0\}$, for all $j \in I \setminus \{i\}\}$, for all $k \geq 1$.*

Intuitively, the event of player i satisfying common structural belief in rationality describes the situation in which i initially as well as at each of his decision nodes, believes that his opponents initially choose rationally, i.e. optimal everywhere in the game tree, initially as well as at each of his decision nodes, believes that his opponents initially as well as at each of their decision nodes believe that their opponents initially choose rationally i.e. optimal everywhere in the game tree, etc. In other words, player i always believes that his opponents choose optimal initial strategies, always believes that every opponent always believes that every other player always chooses an optimal initial strategy, etc. Observe that due to our weaker notion of rationality in Definition 7, it is always possible to define common structural belief in rationality in our epistemic model, contrary to impossibility results, such as by Reny (1992) and Reny (1993), concerning epistemic models with standard rationality. While it is usually not distinguished between initial and actual choice, common structural belief in rationality cannot be generally defined in standard epistemic structures. However, our model is capable of admitting that a player believes that an opponent initially chooses rationally, while at the same time entertaining the belief that the same opponent will actually choose irrationally at some points in the game and thus not carry out the rational strategy of his respective reasoning agent. In our model, a player can reason about both the reasoning as well as the actual play of his opponents. In this context note that Perea (2008) provides an epistemic model in which common structural belief in standard rationality is generally made possible by allowing a player to revise his beliefs about his opponents' utilities during the game, while assuming the respective player's utilities to be constant. As an illustration of the permanent feasibility of common structural belief in rationality in our framework,

consider the dynamic game given in Figure 5.1. Suppose satisfying common structural belief in rationality, *Bob* believes at the beginning of the game as well as at his decision node that *Alice* initially rationally chooses strategy af. It is then possible that *Bob* believes at his decision node that *Alice* actually chooses a strategy different from af, i.e. that game agent $Alice_1$ has picked b and game agent $Alice_3$, for instance, will pick e, while still maintaining his belief in $Alice_r$'s rational choice of the initial strategy af.

Further, note that games are epistemically investigated from a perspective which is completely that of a single player. Even nested belief notions are defined from the viewpoint of a specific player. Understanding interactive epistemology as a theory of reasoning prior to choice, this stance seems natural, since any reasoning process takes place entirely within the reasoning individual, represented here by the reasoning agent.

Connectedness is now used to define the nested epistemic notion of forward belief in future-high-connectedness.

Definition 14 (Forward belief in future-high-connectedness). *Let $\mathcal{M}^\Gamma$ be an epistemic model of an extensive form structure Γ with perfect information and $i \in I$ some player. A type $t_i \in T_i$ expresses* forward belief in future-high-connectedness *if $t_i \in FBH_i = \{t_i \in T_i : t_i \in BH_i^k(x)$, for all $k \geq 1$, for all $x \in X_i \cup \{x_0\}\}$, where $BH_i^1(x) = \{t_i \in T_i : t_i \in BH_{i,j}$ for all $j \in I \setminus \{i\}\}$, and $BH_i^{k+1}(x) = \{t_i \in T_i : supp(\beta_i(t_i, x \mid T_j)) \subseteq BH_j^k(x_j)$, for all $j \in I \setminus \{i\}$, for all $x_j \in X_j$ such that x_j follows $x\}$, for all $x \in X_i \cup \{x_0\}$, and for all $k \geq 1$.*

According to forward belief in future-high-connectedness, a player always believes that his opponents' agents are highly connected at all succeeding nodes, that his opponents believe at all succeeding nodes that their opponents-agents are highly connected at all respectively succeeding nodes, etc. Observe that this epistemic condition implies that at any possible situation in the game, the player believes that any opponent agent at a succeeding decision node is highly connected and hence acts in accordance with the respective initial strategy of his player.

Further, note generally that requiring forward belief in some event E is a considerably weaker epistemic condition than common structural belief in E. Accordingly, a theorem only requiring forward belief in some particular event and not common structural belief is strengthened. To see that common structural belief in E is stronger than forward belief in E, consider a decision node x_i succeeding some node x_j. According to the former epistemic condition i believes

at x_i that opponent j believes E at x_j, while the latter concept of forward belief in E does not put any restrictions on what i believes at x_i what j believes at any preceding decision node, in particular not whether j believes E at x_j. Intuitively, the strength of common structural belief derives from the fact that it concerns any decision node, including respectively preceding ones, relative to a given decision node. In contrast, forward belief concerns only succeeding decision nodes, given a particular decision node.

It is now possible to formulate epistemic conditions for backward induction in terms of connectedness. Proofs are given in an appendix to this chapter.

Theorem 15 (Sufficient conditions for backward induction). *Let $\mathcal{M}^\Gamma$ be an epistemic model of a generic extensive form structure Γ with perfect information and $i \in I$ some player. If $t_i \in T_i$ such that $t_i \in R_i \cap CSBR_i \cap FBH_i$, then $\iota(t_i) = b_i$.*

In our enriched epistemic framework, the preceding theorem provides a foundation for backward induction in terms of connectedness. Intuitively, common structural belief in rationality ensures that the respective player always believes that his opponents initially play rationally i.e. their unique backward inductive strategies, while at the same time he also always believes that his opponents' future game agents actually choose accordingly, by forward belief in future-high-connectedness. Then, i initially chooses his unique backward inductive strategy. In fact, any surprise information that might arise during play is explained by low-connectedness of the deviating game agent, maintaining the belief in future-high-connectedness of all succeeding game agents.

When reasoning about his opponents in the reasoning stage, a player's reasoning agent contemplates both about his opponents' reasoning as well as their actual choices. In fact, it is his conclusion on his opponents' actual choices that finally matters for the decision problem of the player's reasoning agent on the basis of which he then chooses an initial strategy. Conceptually, a type furnished by an epistemic model captures the complete reasoning of the respective player. Indeed, the epistemic states and the reasoning of a player coincide. During the play stage, agents then pick actual choices according to which the dynamic game unfolds. Importantly, actual decisions need not to be in accordance with the underlying reasoning. For instance, a player might change with regards to his underlying connectedness. These interpretative issues will be addressed in Section 5.5.

An epistemic model only prescribes a player's beliefs and intentions, i.e encompasses his reasoning, but it does not prescribe actual choices. Here, our frame-

work precisely captures this basic idea of an epistemic model by distinguishing between initial and actual strategy choice by a reasoning agent and a set of game agents, respectively, and explicitly endorses the possibility of change in a player's decisions. A player's decision furnished by our epistemic model is accurately his initial strategy. Hence, the epistemic foundation for backward induction provided by the above theorem does concern a player's initial and not his actual strategy. However, our framework also permits the formulation of sufficient conditions for backward induction in terms of actual choice as follows.

Corollary 16 (Backward induction play). *Let $\mathcal{M}^\Gamma$ be an epistemic model of a generic extensive form structure Γ with perfect information. If $c_i(\alpha^i_{x_i}, s_i^\alpha \mid t_i) = high$, for all $x_i \in X_i$, and for all $i \in I$, as well as $t_i \in T_i$ such that $t_i \in R_i \cap CSBR_i \cap FBH_i$ for all $i \in I$, then $s^\alpha = b$.*

Accordingly, the backward inductive strategy profile will be played if each player's reasoning agent is rational, expresses common structural belief in rationality as well as forward belief in future-high-connectedness, and each player's game agents are highly connected i.e. actually do carry out their reasoning agent's initial strategy. Again note that actual choice is a property of game agents not of the reasoning agent i.e. the type.

It is now shown that the restrictions imposed on a player's belief revision in line with our epistemic conditions for backward induction do not lead to any contradictions and are indeed always possible.

Theorem 17 (Existence). *Let Γ be a generic extensive form structure with perfect information. Then, there exists an extended epistemic model $\mathcal{M}^\Gamma$ of Γ such that $t_i \in R_i \cap CSBR_i \cap FBH_i$ for all $t_i \in T_i$ and for all $i \in I$.*

As an illustration of this epistemic foundation of backward induction, consider *Bob*'s reasoning in the dynamic game given in Figure 5.1. In order to choose his initial strategy in the reasoning stage, Bob_r hypothetically considers his game agent Bob_2. By forward belief in future-high-connectedness, Bob_r believes at Bob_2 that $Alice_3$ will be high-connected and thus play in line with the initial strategy of her reasoning agent $Alice_r$. Since, by common structural belief in rationality, he believes $Alice_r$ to choose rationally, Bob_r believes at Bob_2 that $Alice_r$ initially chooses f at her final decision node. Therefore, Bob_r believes at Bob_2 that the high-connected $Alice_3$ complies with the initial rational strategy

and thus picks f. Hence, Bob_r initially chooses his rational strategy d, as well as actually in case of Bob_2 being high-connected.

Further observe that our theorem makes explicit a strong principle of epistemic independence needed for backward induction: the observation of a deviating opponent game agent has no influence whatsoever on a player's beliefs concerning any game agents at succeeding decision nodes, who are still believed to be highly connected each. Also note that only requiring forward belief in future-high-connectedness instead of the stronger condition of common structural belief in future-high-connectedness strengthens our epistemic characterisation of backward induction. The epistemic foundation for backward induction in terms of connectedness provided here is interpreted and discussed in Section 5.5.

5.5 Discussion

5.5.1 Dynamics

Our extended epistemic framework, which understands players as sets of agents and models their connectedness, is capable of shedding light on the *dynamic* character of dynamic games.

In a general sense, our framework displays the complete sequential structure underlying the standard extensive form model of dynamic games. According to our framework, a dynamic game has at least three distinguishable stages: the set-up stage, the reasoning stage and the play stage. It is thus made explicit that different agents of a player find themselves in distinct, sequential situations, such as utility assignments before play, reasoning before play and then play at different decision nodes. Moreover, explicating the sequential structure of dynamic games within our framework reveals stability assumptions implicit in the standard extensive form model. The ingredients of the game, including the fact that utilities are determined prior to reasoning and actual play as well as that pre-play strategy choice resulting from reasoning, are supposed to remain invariant during the whole dynamic strategic interaction. Concerning utilities, the assumption of stable preferences of all agents throughout reasoning and play is made explicit by the fact that they respond to the same utility function. Concerning reasoning, our model clarifies that a game agent is presumed at his decision node to comply with his player's instructions i.e. to act in line with the respective initial strategy.

The assumed stability of dynamic games implicit in the standard extensive

form model can be argued to be in tension with its inherent sequential nature. While the latter suggests the possibility of change in dynamic games, the former does not offer enough structure to account for any such changes. This problematic aspect of sequential stability is made explicit and can be relaxed in our framework. In particular, the notion of high-connectedness can be formally introduced which captures the sequential stability of game agents. Intuitively, high-connectedness captures the idea that game agents make choices according to the pre-play instructions of their reasoning agents. Also, the above theorems relate high-connectedness to backward induction reasoning. Note that high-connectedness is a purely behavioural assumption that can be dropped locally, in order, for instance, to account for surprise information in backward induction reasoning. More generally, connectedness can be used to formulate hypothetical reasoning patterns related to the sequential stability of a player and his game agents, which in turn can be applied to epistemic characterisations of game-theoretic solution concepts.

Moreover, by clarifying the sequential character of dynamic games, unveiling stability assumptions and modelling reasoning about connectedness of agents, our framework provides foundations for a realistic interpretation of dynamic games as formal representations of strategic interaction over time. More specifically, two interpretative directions can be taken. Firstly, the very sequential structure of dynamic games as rendered transparent in our framework can be interpreted as temporal. Secondly, the player which is defined as a set of agents with specific tasks in our framework can be interpreted as a person. In particular, interpretations of players in dynamic games can be linked to multiple-self models of personal identity over time by understanding players as multiple-selves and agents as selves. The idea of decision-makers as multiple-selves can thus be addressed in our account of dynamic games. Indeed, the subsequent section interprets dynamic games from a multiple-selves point of view.

5.5.2 Multiple-Self

The conception of player as a set of connected agents can naturally be linked to the notion of connected selves in a multiple-self. Indeed, this idea of understanding agents of players as the different selves of multiple-selves has been employed in the context of extensive form models with imperfect information, such as that of Piccione and Rubinstein (1997). Also, the idea that a player consists of different

acting selves appears in Selten (1975) and Halpern (2001) within the context of the agent normal form. Yet such appearances of the multiple-self concept in game theory lack philosophical foundations. Here, we propose to interpret the notion of player as a multiple-self using theories of personal identity over time, in order to give specific meaning to change of players over time and to the reasoning of players about possible or observed changes of their opponents.

Multiple-self models of personal identity over time can be related to the extended epistemic framework by interpreting the agents as selves and the player as a multiple-self. The purely behavioural notion of connectedness as measuring compliance or deviation of a game agent with or from the initial strategy of the reasoning agent can then be explained by *underlying connectedness* which describes the degree of connectedness in the multiple-self. Substantive interpretations of intrapersonal connectedness, such as with psychological, empathy and memory connectedness render the description of decision-makers in dynamic games more realistic. Adopting such a multiple-self model of personal identity over time, three substantive interpretations of underlying connectedness are now considered and linked to our extended epistemic model and sufficient conditions for backward induction.

Psychological connectedness permits the interpretation of players as consisting of agent-selves with possibly different preferences. That is to say, a supposedly irrational move at some decision node can be interpreted as resulting from different preferences of the respective agent. As an illustration consider the game given in Figure 5.1 and suppose that *Bob* believes $Alice_1$ to rationally play a. Upon observing a surprising move b by $Alice_1$, he can make sense of the low behavioural connectedness of $Alice_1$ as follows: Bob adopts the belief that $Alice_1$ has exhibited deviating preferences from her player. In particular, it is natural to depict a breakdown in psychological connectedness between $Alice_1$ and $Alice_s$. However, it could also be the case that $Alice_1$ has re-evaluated outcomes at later terminal nodes rendering her preferences different from $Alice_3$. Note that a particular contemplation about what precisely has prompted the preference change or about what precisely it consists in, is not formally needed to obtain our sufficient conditions for backward induction in terms of connectedness. Yet, in order to further describe backward induction reasoning, it is possible to provide such more realistic interpretations when viewing a player as a multiple-self. In order to obtain backward induction, an agent who observes a supposedly irrational move

can hence maintain belief in the rationality of the respective opponent, as well as belief in future-high-connectedness by revising his belief in the high psychological connectedness of the deviating game agent. Such belief revision only commits the reasoner to believing there to have been a relevant preference change such as to prompt a local re-evaluation of the payoffs which, in turn, has led to a local deviation from the initial strategy. Note that similarly, Perea (2008) provides an epistemic model for dynamic games in which the possibility of belief change about opponents' utilities during the game is explicitly endorsed, modelled and sufficient conditions for backward induction are derived.

Along similar lines, empathetic connectedness permits us to interpret a supposedly irrational move at an earlier decision node as a local breakdown in the opponent's empathetic connectedness. By ascribing a low empathetic connectedness to some self of another person, it is reasonable to still grant full rationality and reasoning capacity to the remaining game agents and the reasoning agent of that person. Similar interpretations for the dynamic game given in Figure 5.1 as proposed for psychological connectedness are hence available.

With (reductive or non-reductive) memory connectedness, a supposedly irrational move at an earlier decision node can be interpreted as a breakdown of memory between the deviating agent and his player, in particular, that the deviating agent has forgotten the initial strategy. Note that memory connectedness could also be used to interpret issues raised in Piccione and Rubinstein (1997) related to imperfect recall. By assigning a low access to earlier experiences of the deviating agent, a reasoner can revise his belief in the stability of memory of an opponent, excluding the agent from the opponent's agents that share memories while still maintaining belief in rationality. Similar interpretations for the game given in Figure 5.1 as proposed for psychological and empathy connectedness are thus available.

The different interpretations of intrapersonal connectedness in the multiple-self allow both a more fine-grained discussion of reasoning about opponents as well as a more specific interpretation of how observed surprise moves at preceding decision nodes can be explained. When discussing backward induction paradoxes later in this section, we will use a more formalised model of underlying connectedness.

5.5.3 Epistemic Independence

The extended epistemic framework proposed here allows us to give epistemic foundations of backward induction in terms of agent connectedness. This notion of connectedness measures the extent to which game agents are sequentially stable relative to the reasoning agent of the player. In the above theorem, backward induction is assured by explaining surprise information in terms of low-connectedness of the deviating game agent.

Connectedness reflects the fundamental principle underlying any foundation of backward induction. This so-called principle of epistemic independence, which is conceptually discussed by Stalnaker (1998), requires that a player treats any information obtained during the game, such as observed opponents' moves, as irrelevant to his beliefs about opponents' behaviour at later points in the game. This property is at work in our theorem: the observation of a surprising move of an opponent's game agent does not affect a player's beliefs on the behaviour of the respective opponent's future game agents, but rather the concerned game agent is concluded to be low-connected. In other words, his comportment is regarded as isolated and irrelevant to future behaviour of the represented player.

More specifically, forward belief in future-high-connectedness yields the condition of epistemic independence that is implicit in any characterisation of backward induction. At any decision node, forward belief in future-high-connectedness ensures the stability of all game agents at all succeeding decision nodes even if game agents at preceding decision nodes have been deviating from the initial strategy of their reasoning agent. Note that in our framework surprise information precisely consists in deviation from the initial strategy. Epistemic independence is assured by forward belief in future-high-connectedness which, in turn, leads to a behavioural isolation of any surprise information.

In a general sense, note that there is a tension between the sequential stability implicitly assumed to underlie standard accounts of dynamic games and some local breakdown which is needed for epistemic independence. This tension needs to be accounted for in any epistemic characterisation of backward induction. In our framework, the notion of connectedness is used to describe this tension: low-connectedness makes explicit the idea that the sequential stability of the initial strategy can break down locally, while forward belief in future-high-connectedness ensures that the effects of such a breakdown indeed remain local. Connectedness thus makes explicit the crucial rigidity with which epistemic independence re-

quires local breakdowns of sequential stability to be treated.

The multiple-self interpretation introduced above yields further insights into the fundamental principle of epistemic independence. In dynamic games, it is natural to depict a player as a multiple-self whose selves are highly connected according to several interpretations of underlying connectedness, i.e. selves highly connected in terms of psychological features, memory and empathy. Upon receiving surprise information, belief revision according to forward belief in future-high-connectedness sets the behavioural connectedness of the deviating opponent game agent to *low*. It is then plausible to claim that this low behavioural connectedness stems from some breakdown in underlying connectedness. However, forward belief in future-high-connectedness also ensures that any succeeding game agents of the respective opponent are assumed to be entirely unaffected by the deviating behaviour of the particular preceding supposedly low-connected agent under *any* interpretation of underlying connectedness, such as psychological, empathy and memory connectedness. Hence, forward belief in future-high-connectedness reveals that foundations for backward induction have commonly been tacitly assuming a much stronger epistemic independence assumption. Indeed, it is plausible to require forward belief in future-high-connectedness for any underlying connectedness as well. Note that assuming such epistemic independence with regards to any underlying connectedness is considerably strong, which suggests that the assumption of epistemic independence is much stronger than commonly assumed.

Furthermore, our framework is capable of clarifying the epistemic conditions for backward induction in Aumann (1995). In his framework, Aumann uses an entirely static epistemic operator that refers to the beginning of the game. Once fixed, the epistemic state of a player concerns a single point in time and does not change. It is hence difficult to account for belief revision in this framework. However, Aumann's key nested epistemic notion of common knowledge of rationality can be interpreted as being equivalent to our concept of common structural belief in rationality. Indeed, rationality refers to a player's initial strategy fixed before the game and rigidity of a belief in a rational initial strategy is thus possible to entertain in our model. However, the belief in the actual choice of the opponents may change at different points in the game. It can hence be claimed that Aumann implicitly endorses some kind of high-connectedness assumption, requiring it to be common knowledge that a player never actually changes his intended initial

strategy. This implicit assumption is explicated by our forward belief in future-high-connectedness condition. By understanding strategies as intentions or, more precisely, initial strategies, Aumann is able to obtain backward induction with an entirely static epistemic operator.

5.5.4 Backward Induction Paradoxes

The so-called backward induction paradoxes have been addressed by, for instance, Selten (1978), Rosenthal (1981) as well as Binmore (1987), and identify games in which backward inductive reasoning leads to rather implausible and counter intuitive strategy choices. In this context, a crucial argument against the plausibility of backward induction criticises that the reasoning method does not take into account any observed past behaviour at all, even when the backward inductive strategy profile is contradicted during actual play. In fact, our framework can be used to juxtapose belief revision patterns in line with such a plausibility requirement, and to contrast them to belief revision policies sufficient for backward induction reasoning according to the above theorem. Recall that the latter belief revision policies require a player to set the connectedness of a deviating agent to *low* and to maintain belief in the high-connectedness of each of the opponent's future agents. In contrast, belief revision policies in line with the plausibility requirement set the connectedness of all future game agents of the relevant opponent to *low* upon observing an opponent game agent deviate. Thus, the intuition is captured that the respective game agents actually play a strategy different from the initial strategy believed to be chosen by their reasoning agent.

As an illustration of this comparison between these two kinds of belief revision policies for dynamic games with perfect information, consider the dynamic game given in Figure 5.1. Suppose *Bob* reasons in line with the conditions of Theorem 15 and hence in line with backward induction. In case of him surprisingly observing $Alice_1$ to choose b, he sets her connectedness to *Alice*'s reasoning agent to *low*, while keeping his belief in the high-connectedness of *Alice*'s future game agent $Alice_3$. Alternatively, suppose now that *Bob* when observing $Alice_1$'s deviating move still believes that *Alice*'s reasoning agent has chosen the backward inductive strategy as initial strategy, but that the game agents play a strategy different from the initial strategy. He thus also sets the connectedness between $Alice_3$ and *Alice*'s reasoning agent to *low*. Note that he is free to believe what $Alice_3$ will choose. For instance, if he believes that she will pick e, then he can

only optimally select c at his decision node.

We now focus on the Rosenthal (1981) approach to the backward induction paradoxes. On his account, small probabilities of future deviating moves are introduced into dynamic games and interpreted as the players' intersubjective beliefs about future moves of opponents. In our framework, such probabilities can be elucidated by introducing a more fine-grained belief revision on deviating moves. More precisely, we introduce probabilistic beliefs on how likely it is that opponent game agents deviate from their respective initial strategy. These probabilistic beliefs are updated by a player's belief about the underlying connectedness of an opponent, which in turn depends on his beliefs on the opponent's behavioural connectedness. Recall that underlying connectedness describes the degree of connectedness of psychological features, memory and empathy in a multiple-self.

A player could form probabilistic beliefs on future deviating moves of an opponent as follows. Firstly, suppose a player observes a move by an opponent agent that deviated from his respective initial strategy and explains it with the latter's low *behavioural connectedness*. Note that this low behavioural connectedness can be treated as information about the opponent. Secondly, suppose further that a player has a belief about the *underlying connectedness* of the opponent's person. It is then natural to update these beliefs with the behavioural observation. In other words, players can learn about their opponent's character during the game. Thirdly, also suppose that a player entertains beliefs about his opponents' future behaviour. Then, it seems reasonable to update the latter beliefs with his beliefs about the respective opponent's underlying connectedness.

Let a specific underlying connectedness be assigned to any *pair* of agents in a player: $c : i \times i \to \{high, low\}$, where a player i is conceived of as a set of agents according to Definition 3. Such an underlying connectedness function can be suitably interpreted as an opponent's belief about the connectedness between any pair of agents in a player. For instance, within the context of the game given in Figure 5.1, starting with a natural belief in high-connectedness from sequential stability as well as a belief that *Alice*'s initial strategy is the backward inductive one, upon observing that $Alice_1$ chooses b, backward induction can only be behaviourally maintained for *Bob* by updating his belief about *Alice*'s underlying connectedness function as follows: set the connectedness of any pair of *Alice*'s agents involving $Alice_1$ to *low*, i.e. $c(Alice_s, Alice_1) = low$, $c(Alice_r, Alice_1) = low$,

$c(Alice_1, Alice_3) = low$, $c(Alice_1, Alice_s) = low$, $c(Alice_1, Alice_r) = low$, and $c(Alice_3, Alice_1) = low$, while maintaining the belief in the high connectedness between any two other agents, i.e. $c(Alice_s, Alice_r) = high$, $c(Alice_s, Alice_3) = high$, $c(Alice_r, Alice_3) = high$, $c(Alice_r, Alice_s) = high$, $c(Alice_3, Alice_s) = high$, and $c(Alice_3, Alice_r) = high$. Clearly, such a belief revision policy is implausible: believing that $Alice_1$ is low-connected to all other agents while still preserving belief in the high-connectedness between all other pairs of agents seems to completely deny any relevance of $Alice_1$ with regards to *Alice* as a multiple-self.

Now consider a more fine-grained underlying connectedness function, as introduced Chapter 2, which expresses degrees of connectedness of pairs of agents, namely $c : i \times i \rightarrow [0, 1]$. Then, according to psychological connectedness, the degree of connectedness then measures the similarity of preferences between agent-selves. Similarly, memory and empathy connectedness can be seen as a matter of degrees rather than binary. Using such a more fine-grained underlying connectedness function, the following belief revision upon receiving surprise information seems plausible. The supposedly low behavioural connectedness of the deviating game agent $Alice_1$ induces *Bob* to believe that there is some agent of *Alice* to which $Alice_1$'s connectedness is strictly less than 1. In other words, some failure of the underlying connectedness has to be assumed in order to explain the low behavioural connectedness. Then, such revised beliefs in an underlying connectedness function can be used to update beliefs about future moves of an opponent. More specifically, underlying connectedness can determine a player's probabilistic beliefs about how likely an opponent game agent will deviate from the initial strategy at future nodes. Conditionalising on the fact there is some agent of *Alice* to which $Alice_1$'s connectedness is strictly less than 1, plausible updating rules render *Bob*'s probabilistic beliefs in future deviation of any of *Alice*'s game agents strictly positive, since each agent has some relevance to his respective player as a multiple-self. Intuitively, upon believing that there is at least some agent-pair in an opponent which is not perfectly connected, the respective player's belief about the future deviation of his opponent's game agents will be strictly positive, as future game agents may also be disposed to deviate from the reasoning agent's initial strategy as exhibited by the particular game agent that has already deviated.

Further plausible constraints on such updating patterns can be introduced. For instance, it is possible that under the interpretation of psychological connec-

tedness, the underlying connectedness changes more drastically than under the interpretation of memory connectedness. Consider a preference change of one game agent. If all other agents' preferences remain stable, then there will be a low-connectedness between the respective game agent and all other agents. However, in the context of memory connectedness, it could be the case that one game agent has forgotten the initial strategy, while all other agents still remember the initial strategy well. Therefore, it is at least plausible to require monotonicity to be respected in updating beliefs about future deviation for psychological connectedness. With different interpretations, the breakdown of connectedness can be more or less wide in scope in terms of how many agents are affected. Which or whether all of these interpretations are endorsed depends on how realistic a model of the decision-maker is intended. Naturally, it is beyond the scope of the present work to explore plausible updating rules and interpretations in greater detail.

More generally, our framework permits us to argue that backward induction reasoning is implausible when underlying connectedness is interpreted as a belief about the opponent's character and probabilistic beliefs about future deviation of opponents are updated on the basis of beliefs about underlying connectedness.

5.5.5 Trembling Hand

It is possible to interpret the idea of a perturbed game in Selten (1975) with connectedness as understood in our framework. The claim that a deviating move is due to the respective player exhibiting a 'trembling hand', and thus making a slight mistake by picking an irrational action with small probability, can be expressed and explained in our model. The agents of a player are assumed to be highly connected and play in line with the initial strategy with almost certainty, yet with small probability their underlying connectedness is low and subsequent behaviour deviates. In other words, a given game agent might – despite it being assumed to be very unlikely – tremble in implementing the intended initial strategy of his player's reasoning agent. Such trembles can be used as explanations for observed deviations from an opponent's supposed initial strategy in belief revision policies. More precisely, whenever a player who believes in the high-connectedness of the game agents of each of his opponents is surprised by a move of a game agent which contradicts the initial strategy he believes the respective opponent's reasoning agent to have chosen, he can then separate that

game agent. Indeed he can only assign low-connectedness to that particular game agent, while keeping fixed the high-connectedness of the respective opponent's other game agents. Such isolated behaviour of this given deviating game agent is explained as a mistake on the agent's part.

Note that such a specific trembling hand vindication of deviating behaviour corresponds to a particular belief concerning the underlying connectedness of the relevant game agent. Intuitively, the trembling hand is a physical metaphor for the failure to complete a given task, despite having had appropriate dispositions to perform it. In terms of underlying connectedness, it is possible to interpret a trembling hand with low empathy connectedness, while at the same time psychological as well as memory connectedness are high. Hence, the deviating game agent is now supposed to be lowly connected to the player as a multiple-self: even though he has the same preferences and perfect memory, he somehow slips and makes a mistake. Note that such belief revision policies are close to the ones sufficient for backward induction and far from the supposedly more plausible ones with regards to the backward induction paradoxes in terms of their general intuition. Lexicographically speaking, whenever a surprise move of some game agent contradicting the supposed initial strategy of his respective reasoning agent is observed, the state in which the deviating agent is lowly connected and others are highly connected is deemed infinitely more likely than the state where the player's future agents are also lowly connected to the respective player. The key to the construction and comparisons of such belief patterns is our notion of initial strategy, which can be contrasted with the same player's actual strategy, and hence belief about initial choice can be juxtaposed with belief about actual choice.

As an illustration of the idea of a trembling hand in the context of our framework, consider the dynamic game given in Figure 5.1. Suppose *Bob* initially believes all game agents of *Alice* to be high-connected and at Bob_2 that *Alice*'s initial strategy is backward inductive one af. However, at Bob_2 he then has has to accommodate the surprise information that $Alice_1$ has actually chosen b. Explaining this deviating behaviour with an exceptional mistake incurred by $Alice_1$ in implementing $Alice_r$'s plan, *Bob* sets the connectedness of game agent $Alice_1$ to low, yet preserves his belief in the high-connectedness of *Alice*'s future game agent $Alice_3$. His unique optimal choice is hence given by d.

5.6 Conclusion

By rendering transparent relevant yet usually neglected processes linked to dynamic games we clarify their inherent dynamics. We analyse the sequential structure of dynamic games with a three-stage account, which defines player and strategy relative to these dynamic stages. A sequential stability assumption underlying the standard extensive form model of dynamic games is made explicit in our account. To describe reasoning in dynamic games, a more general epistemic model is proposed that is capable of formalising the notion of agent connectedness. Such an enriched framework sheds light on backward induction reasoning. Formally, we provide sufficient conditions for backward induction in terms of connectedness, as well as an existence result ensuring that our conditions are indeed possible. Conceptually, the essence of backward induction can be explicated, since surprise information is explained with low-connectedness of the deviating agent. Also, the epistemic independence assumption underlying any foundation of backward induction can be shown to be considerably stronger than usually assumed. Our framework makes explicit that any underlying connectedness of players as multiple-selves has tacitly been assumed to be high.

In a general sense, our frameworks provides adequate foundations for interpreting the sequential structure of dynamic games in temporal terms. In particular, defining a player as a set of agents enables a more realistic interpretation of decision-makers in dynamic games. Using the multiple-self model of personal identity over time also provides richer descriptions of players, for instance, with regards to psychological, empathy and memory connectedness. Hence, our framework is especially relevant for economics and the social sciences, where players should be interpreted as persons existing over time.

Finally, the framework proposed here could also be employed to shed light on the sequential structure and dynamics of games of imperfect information as well as to to clarify corresponding reasoning and solution concepts. It would be of particular interest to search for sequential stability requirements for forward induction reasoning in terms of agent connectedness. Intuitively, actual choice of a game agent should then be believed to be highly relevant to actual choice of the respective player's future game agents.

5.7 Appendix: Proofs

Proof of Theorem 15

Proof. Suppose that $t_i \in T_i$ such that $t_i \in R_i \cap CSBR_i \cap FBH_i$. We show that t_i initially assigns the backward inductive action to each decision node $x_i \in X_i$, i.e. $\iota(t_i) = b_i$. Consider a decision node $x_i \in X_i$ of player i. Suppose that x_i is an ultimate decision node. Then, by rationality of t_i, $\iota_i^{t_i}(x_i) = b_i(x_i)$. Suppose that x_i is a pre-ultimate decision node. Since $t_i \in FBH_i$, type t_i believes at x_i that every opponent game agent $\alpha_{x_j}^j$ is high-connected to his respective player j and hence chooses according to j's initial strategy at every ultimate decision node x_j succeeding x_i. As $t_i \in CSBR_i$, he also believes at x_i in j's rationality i.e. that j's initial strategy is rational. Hence, t_i believes that every high-connected opponent game agent $\alpha_{x_j}^j$ does indeed choose rationally at every x_j succeeding x_i, and thus picks the unique backward inductive action $b_j(x_j)$ there. Therefore, the unique optimal action for i at x_i is the backward inductive one and rationality of t_i ensures that $\iota_i^{t_i}(x_i) = b_i(x_i)$. Now suppose that x_i is a pre-pre-ultimate decision node. Since $t_i \in FBH_i$, type t_i believes at x_i that every opponent game agent $\alpha_{x_j}^j$ is high-connected to his respective player j and hence chooses according to j's initial strategy at every decision node x_j succeeding x_i. Note that every opponent decision node x_j succeeding x_i is either pre-ultimate or ultimate. Suppose that x_j is ultimate. As $t_i \in CSBR_i \cap FBH_i$, type t_i believes at x_i in j's rationality i.e. that j initially chooses rationally, as well as that every high-connected opponent game agent $\alpha_{x_j}^j$ does indeed choose rationally at every ultimate decision node x_j, and thus picks the unique backward inductive action $b_j(x_j)$ there. Suppose that x_j is pre-ultimate. Since $t_i \in FBH_i$, type t_i believes at x_i that at any immediately succeeding opponent decision node x_j, the respective opponent j believes that his opponents' game agents are high-connected, and thus act in accordance with their respective player's initial strategy, at all succeeding ultimate decision nodes. Also, by $t_i \in CSBR_i$, type t_i believes at x_i that his opponents believe at all succeeding nodes in their opponents' rationality i.e. that their opponents have initially chosen rationally. Hence, t_i believes at x_i that at any immediately succeeding opponent decision node x_j, the respective opponent j believes that his opponents' high-connected game agents play rationally at every ultimate decision node succeeding x_j. Moreover, by $t_i \in CSBR_i$, type t_i also believes at x_i in j's rationality, i.e. in a rational initial strategy choice of j. Since $t_i \in FBH_i$, it then follows that

he believes at x_i that every high-connected game agent $\alpha^j_{x_j}$ does indeed choose rationally at the respective pre-ultimate decision node x_j. But as t_i believes at x_i that j believes at x_j that j's opponents choose the backward inductive actions at all ultimate decision nodes succeeding x_j, in fact t_i believes at x_i that $\alpha^j_{x_j}$ picks his unique backward inductive action $b_j(x_j)$ at x_j. Therefore, since t_i believes at x_i that at any succeeding decision node the respective opponent game agent chooses the backward inductive action, the unique optimal action for i himself at x_i is the backward inductive one and $\iota_i^{t_i}(x_i) = b_i(x_i)$ obtains by rationality of t_i. By induction, it follows that at any $x_i \in X_i$, type t_i believes that his opponent game agents choose the unique backward inductive action at any x_j succeeding x_i, and hence, being rational, t_i initially assigns the unique backward inductive choice to each of his decision nodes, i.e. $\iota_i(t_i) = b_i$. □

Proof of Corollary 16

Proof. Consider $i \in I$ and suppose that $t_i \in T_i$ such that $t_i \in R_i \cap CSBR_i \cap FBH_i$. It follows from Theorem 15 that $\iota_i^{t_i}(x_i) = b_i(x_i)$ for all $x_i \in X_i$. Since $c_i(\alpha^i_{x_i}, s_i^\alpha \mid t_i) = \text{high}$, for all $x_i \in X_i$, each high-connected game agent of player i will indeed choose the backward inductive action $s_i^\alpha(x_i) = b_i(x_i)$ at any $x_i \in X_i$, respectively. Therefore, i's actual backward inductive strategy choice $s_i^\alpha = b_i$ obtains. □

Proof of Theorem 17

Proof. Consider some player $i \in I$ and for every opponent $j \in I \setminus \{j\}$ and node $x \in X_i \cup \{x_0\}$, let $b_j^*(x)$ be the strategy that prescribes the unique action on the path to x at every node $x' \in X_j$ preceding x, and that prescribes the unique backward inductive action at every node $x' \in X_j$ not preceding x. Fix type spaces $T_i = \{t_i\}$ for every player $i \in I$ such that $\iota_i(t_i) = b_i$ and $supp(\beta_i(t_i, x)) = \times_{j \in I \setminus \{i\}}(\{b_j^*(x)\} \times \{t_j\})$ for all $x \in (X_i \cup \{x_0\})$. Consider $i \in I$ and observe that t_i believes at every point in the game that his opponents initially as well as at all succeeding nodes actually choose their backward inductive actions. Since i has been arbitrarily picked, every player's type believes throughout the game that his opponents initially as well as at all succeeding nodes actually choose their backward inductive actions. In particular, it thus follows that $t_i \in FBH_i$. Besides, note that $t_i \in R_i$, since $\iota_i(t_i) = b_i$, which maximizes conditional expected utility for his conditional beliefs which always assign probability 1 to his opponents' future play being in line with backward induction. Since i has

been arbitrarily picked, $t_i \in R_i$ obtains for all $i \in I$. As every type at every point in the game only deems possible opponents' types that are rational, it follows in particular that $t_i \in CSBR_i$. □

Chapter 6

Preference Change

Summary. This chapter analyses temporal dynamics and gives an account of dynamic inconsistency. Two families of approaches to dynamic inconsistency are identified: firstly, those that use hyperbolic discounting functions to describe dynamically inconsistent decision-makers as myopic, and secondly, those that postulate multi-selves models that capture different motivations and time horizons which can lead a decision-maker to (fail to) control himself in the face of temptation. In order to achieve a simpler characterisation of dynamic inconsistency, we reconsider both hyperbolic discounting and multi-selves models in the more general model of connectedness in the multiple-self. A simple specification of this model can motivate hyperbolic discounting, and an extended version of it can be used to reformulate the multi-selves models, using a less complex structure that can be better motivated. Moreover, the latter allows us to distinguish between conflicts in connectedness and conflicts in goodness evaluation.

6.1 Introduction

Applications of decision theories usually assume that decision-makers have stable preferences. Yet, the preferences of real-world decision-makers often change over time: a gourmet might not want to eat at his favourite restaurant any more after learning that it hired a new chef, college graduates make career choices they would not have made before their education, and adolescents sneer at the music they listened to when they were children. Such changes in preferences can have diverse reasons, as these examples suggest: they can stem from changes in beliefs, for example through acquiring information or learning, or from changes in tastes,

for example through habituation. As reviewed in Chapter 2, such changes in preference can be modelled by theories of Bayesian conditionalisation as well as newer approaches that deal with changes in taste.

Changes in preference can also occur in less laudable circumstances, such as when individuals appear to contradict themselves over time: for instance, consider someone who plans to eat a salad for dinner as part of a healthy diet only to choose steak when ordering and then again trying and failing in adhering to a healthy diet later on, or someone joining a gym, stopping to go after a while and then renewing the membership at the next opportunity without going more often later on, or someone making an attempt to quit smoking, succumbing to the habit soon after and then making a new attempt, and so on. These examples suggest that decision-makers can have conflicting preferences. To be sure, weighing up different desires is part of any decision-making process. It does not seem to be inherently irrational to have both a desire to eat a steak and a desire to be healthy. Yet, there is still something tragic about the individuals in the above examples, in the sense that they oscillate between fulfilling incompatible desires, thus failing to achieve long-term goals. It is hard to imagine them as satisfied.

The changes in preferences associated with the above examples do not seem to be very well motivated. Indeed, a lot of mitigating details would need to be in place to defend a decision-maker as rational who switches preferences constantly, for example when first salad is preferred to steak (when planning a healthy diet), then steak is preferred to salad (when ordering), and then again salad is preferred to steak (the day after). Yet, individuals are prone to exhibit such patterns of preference change from time to time. Indeed, preference change as described above has been researched extensively in behavioural economics, where it is often described as 'failure to self-control' and more generally labelled as 'dynamic inconsistency'. This literature has collected compelling evidence for the persistence of such phenomena, through experiments as well as empirical studies of consumer behaviour (Loewenstein and Prelec (1992), Loewenstein (1996), Connor *et al.* (2002), Loewenstein and Read (2003)).

The empirical evidence raises the question of how dynamic inconsistency can be modelled. More precisely, how can dynamic inconsistency be predicted, explained, and resolved, i.e. reconciled with standard decision theories? One important modelling device consists in hyperbolic discounting functions (Strotz (1956), Laibson (1997), Frederick *et al.* (2002), Angeletos *et al.* (2001)). Hyper-

bolic discounting functions capture extreme short-sightedness of decision-makers and can therefore model those types of dynamic inconsistency that are due to a strong bias for the present and near future. Another important family of approaches consists in the so-called 'multi-selves' theories (Schelling (1980), Thaler and Shefrin (1981), Schelling (1984), Benabou and Pycia (2002), Fudenberg and Levine (2006), Read (2006)). In these models, a decision-maker is assumed to consist of a far-sighted 'planner'-self and short-sighted 'doer'-selves. The interaction between planner and doers is modelled in extensive-form games and dynamic inconsistency occurs when the doer-selves retain the upper hand in such interactions. This permits us to model a wide range of types of dynamic inconsistency, including cases of lack of self-control in the face of temptation. They also allow us to outline how precommitment strategies can resolve dynamic inconsistency (for instance, when someone instructs a friend before dinner to order salad in order to prevent himself from choosing steak). The multi-selves models are mostly consistent with hyperbolic discounting – in fact, many accounts of both hyperbolic discounting and multi-selves show how the two modes of modelling can be formally equivalent (Fudenberg and Levine (2006), Xue (2008)), and Ainslie (1992, 2001, 2005) combines aspects of hyperbolic discounting and multi-selves in his 'picoeconomics' approach.

This chapter analyses such behavioural economics' accounts of dynamic inconsistency. In a first step, we show in what sense the two types of accounts fullfil the modelling goals of *predicting*, *explaining* and *resolving* dynamic inconsistency. Simply put, hyperbolic discounting offers predictively accurate models of dynamic inconsistency, yet they have explanatory deficiencies, and few resources to resolve dynamic inconsistency. Multi-selves models attempt to elucidate the processes that lead to dynamic inconsistency and its resolution, yet their structure is often complex, and it is unclear how it relates to standard decision-theoretic frameworks.

To improve on these deficiencies, we introduce multiple-self models of personal identity over time, and reconsider the approaches of hyperbolic discounting and multi-selves in this framework. This discussion provides both a more explicit motivation for hyperbolic discounting models, and a simpler structure for the multi-selves models in the literature. The multiple-self models of personal identity over time introduced here thus combine and improve on the virtues of both hyperbolic discounting and multi-selves approaches; in particular, their structure

is simpler, and they make transparent the additional assumptions required in standard-decision theoretic approaches to model dynamic inconsistency.

This chapter proceeds as follows. Section 6.2 introduces two salient examples of dynamic inconsistency, highlighting the problems of present bias and temptations. Section 6.3 critically scrutinises hyperbolic discounting and 'multi-selves' approaches in behavioural economics that attempt to predict, explain, and resolve dynamically inconsistent behaviour. Section 6.4 presents multiple-self models of personal identity and shows how they complement and elucidate existing models. Section 6.5 concludes.

6.2 Preference Change and Dynamic Inconsistency

Changes in preferences can occur in many different circumstances. As mentioned in the introduction, they are often well-motivated, for instance when they reflect some kind of learning. For example, take an agent who prefers going to the beach over staying at home. Suppose she listens to the weather forecast and learns that it is likely that it is going to rain tomorrow. The agent might now prefer staying at home over going to the beach. Such changes in preference that stem from learning can be modelled as cases of Bayesian conditionalisation. In addition to models of learning, there is a recent literature (as reviewed in Chapter 2) that develops more permissive models that allow one to explain preference change by well-motivated taste changes, such as in habituation or when developing refined tastes (such as Bradley (2009a), Dietrich and List (2009)).

Some changes in preference cannot be explained and defended as rational by underlying processes such as learning and habituation. For instance, take a decision-maker who switches back and forth between different preferences over time without any good reason to do so. While not all such changes in preferences are flatly irrational, many of them cannot be particularly well motivated. In the behavioural economics literature, the term 'dynamic inconsistency' denotes such changes in preferences that leads to behaviour which is hard to defend as rational. In the following, we consider two particularly interesting examples of dynamic inconsistency.

Present Bias. Consider an agent who is presented with two choices. Firstly, she is choosing between receiving one apple today and two apples tomorrow. Secondly, she is choosing between receiving one apple in 999 days and

receiving two apples in 1,000 days. For simplicity, suppose that the agent has to make a choice, i.e. declaring her indifference is not an option. Then, four possible combinations of choices are possible: two 'symmetric' ones where she chooses either one apple or two apples in both of the choices, and two 'asymmetric' ones where she chooses differently in the two choices. Real-world agents often choose 'asymmetrically' by choosing to receive one apple today in the first choice and two apples in 1,000 days in the second choice.

Now consider the implications of the four possible choices. It is easy to see that the two symmetric choices are compatible with recommendations from standard decision theories, as either one of those choices can be taken as reflecting preferences that are representable by a utility function. However, the two asymmetric choices are troubling. Consider the choice pattern of receiving one apple today in the first choice and two apples in 1,000 days in the second choice. Since the agent has firstly chosen to receive one apple today over two apples tomorrow, after 999 days she will prefer to receive an apple on that day, rather than waiting for receiving two apples on day 1,000. Yet, this goes against her earlier choice of receiving two apples in 1,000 days rather than one in 999 days. Moreover, 'giving up' on her preference for receiving one apple on day 999 at this point for the sake of avoiding a contradiction does not solve the problem, as it implies this: on day 999, she will now have a preference for receiving two apples tomorrow rather than one today, which goes against her earlier choice. (This problem holds vice versa for the other 'asymmetric' choice.) It seems that by choosing 'asymmetrically' in the above set-up, decision-makers will have inconsistent preferences over time, due to the fact that after waiting 999 days, the second choice becomes equivalent with the first one.[1] Choosing differently when the same options are at stake implies dynamically inconsistent preferences.

Dynamic inconsistency is revealed in the 'asymmetric' choices, which leave the decision-maker with options that she does not prefer at either one of the later days. It seems that an obvious normative recommendation in this type of choice is to demand the decision-maker to settle for one of the 'symmetric' choices. From a descriptive point of view, a prominent explanation for the type of choice cited as

[1]That is, we assume here that the preferences that lead to the initial choices remain fixed, such that choosing one apple today over two apples tomorrow implies that on a different day, there will be a preference for choosing one apple on *that day*. Then, the two choices become equivalent after 999 days and the inherent inconsistency in the initial preferences is exposed.

the popular one in the above example is that it is due to a bias for the present, or near future. This kind of 'myopia' of agents is captured in hyperbolic discounting functions, indeed, the above example is frequently employed in the hyperbolic discounting literature, and has become one of the standard examples to illustrate its mechanism (such as in Laibson (1997)). Dynamic inconsistency of the sort described in the above example will be henceforth called 'present bias', as it seem to primarily arise out of an initial bias for the present, or the near future.

The above case of dynamic inconsistency and its explanation are reasonably straightforward. Yet, there are also cases in which the explanation of why decision-makers exhibit dynamically inconsistent preferences is more involved, such as in the following one.

Temptation. Consider an agent who is about to finish his day's work in the office, and who can choose between going home directly and going to the pub. Once in the pub, he can choose whether to have one drink and go home or to stay longer and get drunk, waking up with a hangover the next day. Suppose that the agent, while in the office, finds that the best option would be to (a) have one drink and go home, the second best option to (b) go home directly, and the third best option to (c) get drunk. Further suppose that the agent finds it highly likely that once in the pub, he will be unable to go home after one drink and rather stay and get drunk. Finally, suppose that the agent also finds it likely that on the next day, he will assess the three possible options in the same way he assesses them when still in the office. Despite all this, anecdotal evidence suggests that in such choices, real-world agents often end up choosing what is ultimately the third best option.

Consider the three possible courses of action in the example. The preferences a over b, and b over c can be thought of as the decision-maker's 'overall' evaluation, as both before and after all consequences have materialised, that is his evaluation. Yet, when attempting to pursue the best option by going into the pub, it so happens that his preferences change momentarily and he prefers c over all other outcomes. This change in preference, however, is reversed again to the overall evaluation by the time the hangover starts. It is hard to defend going to the pub in the pursuit of a in this type of scenario, as the disastrous effects of this choice were well established beforehand. However, this does not prevent real-world agents from making choices of this type. Dynamic inconsistency of this

sort will be called a case of 'temptation' in the remainder of the paper, as it arises out of temptation for an option that is associated with the wrong kind of preference change.

Theoretical approaches in behavioural economics have attempted to model these more complex cases of dynamic inconsistency by envisaging a decision-maker as composed of different 'selves' which compete for influence on the behaviour. Early treatments of such approaches are in Strotz (1956), Peleg and Yaari (1973), Schelling (1980), and Schelling (1984). In those models, different selves embody the diverging evaluations of the decision-maker that lead to dynamically inconsistent preferences: a planner-self assesses the prospect, only to be undermined by a doer-self. Such 'interactions' between selves that are often modelled as dynamic games in the recent literature, which permit us to discuss how both dynamic inconsistency and successful self-control can arise.

The two cases of dynamic inconsistency discussed here, 'present bias' and the 'temptation', share a common structure: they depict changes in preferences that lead to contradictions in behaviour which cannot be easily defended as rational. While different in their degree of complexity, and different in the kinds of explanations that would elucidate them, they pose the problem of how to explain such changes in preference. Many other examples of dynamic inconsistency are variants of those two cases. Consider again the cases of dynamic inconsistency that were briefly mentioned in the introduction, such as unhealthy eating, unsuccessful fitness regime, reluctant addiction, or cases of procrastination. It seems that in most of those cases, both temptation and a bias for the present could play a role in explaining dynamically inconsistent behaviour. Indeed, the theories reviewed in the next section have advocated different, yet ultimately compatible, approaches to do so.

Before reviewing those theories, a few terminological remarks are in order. Dynamic inconsistency as discussed in the above refers to ill-motivated changes in preference that lead to irrational behaviour. Note that in the literature that exclusively deals with a normative assessment of dynamic inconsistency, it is sometimes understood as a property of *choice*, such as in Hammond (1976). The discussions in this literature focus on similar examples as described above. Indeed, the temptation case is a variant of the popular case of 'Ulysses and the Sirens', the potential addict in Hammond (1976), as well as the piano player in Bratman (1996). Furthermore, note that in the behavioural economics lit-

erature, ill-motivated preference change is sometimes labelled more generally as 'preference reversal' which includes ill-motivated preference change due to framing effects, risk preferences, and rules-of-thumb.[2] We will proceed using the term dynamic inconsistency as it specifically refers to the temporal aspect of the changes in preferences. We now turn to review theories that attempt to model dynamically inconsistent decision-makers.

6.3 Theories of Dynamic Inconsistency

This section discusses approaches to modelling dynamic inconsistency that have been put forward in behavioural economics. Two families of approaches, hyperbolic discounting and 'multi-selves' models, are discussed. A number of papers have demonstrated that the hyperbolic discounting and multi-selves approaches are compatible, such as Xue (2008). Indeed, Fudenberg and Levine (2006, 1469) also maintain that their multi-selves model is consistent with quasi-hyperbolic discounting (Laibson, 1997). In reviewing these two families of compatible approaches, we will show that hyperbolic discounting theories have some explanatory deficiencies, and that the 'multi-selves' approaches have a complex structure that is difficult to motivate. This motivates the development of simpler models of dynamic inconsistency, by employing multiple-self models of personal identity over time in the next section.

6.3.1 Hyperbolic Discounting

In hyperbolic discounting theories of dynamic inconsistency, decision-makers are modelled as short-sighted, discounting much more heavily in the short-run than in the long-run. Roughly speaking, any hyperbolic discounting function characterises three time horizons of a decision-maker: (i) the present and immediate, which is given full weight, (ii) the horizon, in which discounting factors are sharply declining between different periods, and (iii) the far future, in which discounting factors are very similar between different periods.

[2] In a strict sense, the 'dynamic' aspect of the inconsistency in the present bias case is by and large implicit. However, note how the inconsistency in this case reveals as the result of time's passage: only on day 999 it becomes apparent that there is an inconsistency. This is different from other cases of preference reversals, such as those observed in the example due to Allais (1953). While it is possible to analyse Allais-type cases as sequential decisions (Steele, 2007), it cannot be said in those cases that the preference reversal is closely connected to the temporal dimension.

As a simple example, take discounting for delay, given by $D(t) = \frac{1}{t}$, where t equals the length of delay (Ainslie, 1975, 1992). In this function, the present and immediate are periods 0 and 1, which are given full weight, i.e. $D(0) = 1$ and $D(1) = 1$. In the horizon, which starts after period 1, the discounting factor is sharply declining (compare, for instance $D(0) = 1$, to $D(2) = .5$ and $D(3) = \frac{1}{3}$). In the far future, the discounting factors are similar between the different periods (such as $D(999) \approx .001$ and $D(1,000) = .001$). As reviewed in Chapter 4, different hyperbolic discounting functions have been proposed that yield slightly different numerical values for the three time horizons.

Capturing the myopia of decision-makers is the key property of hyperbolic discounting in the context of discussing dynamic inconsistency. It is immediately obvious that, for instance, discounting for delay captures present bias. As seen above, the difference in the discounting factors between periods 1 and 2 in the near present is much larger than the difference between periods 999 and 1000, which is negligible. Applying those discounting factors to the example of present bias given earlier, we see how the difference of one period has a much larger impact in the short run than in the long run, leading to the preference of an immediate apple over two apples in the horizon, and preferring two apples over one apple in the far future. Once the far future becomes the present, the dynamic inconsistency is revealed. Other functions, such as generalised hyperbolic discounting and quasi-hyperbolic discounting give similar (and in many contexts slightly more accurate) results. The diversity of the hyperbolic discounting proposals is due to their quintessentially descriptive nature: their primary role is to capture the myopia or short-sightedness of agents in many different circumstances, in order to include more precise characterisations of attitudes to intertemporal prospects in consumer models. As such, hyperbolic discounting theories lend themselves to modelling dynamic inconsistency. As an example for applications for such models, consider consumer behaviour in the fitness and dieting industry where dynamic inconsistency as suggested by the earlier examples is persistent and widespread. This suggests that such straightforward examples of present bias can be modelled adequately by hyperbolic discounting functions.

Can hyperbolic discounting functions also model the more complex cases of temptation? Applying hyperbolic discounting to the pub example suggests that correct predictions can be made in such cases: a short-sighted decision-maker will give full weight to the immediate benefits of going to the pub and heavily

discount next day's after effects, while at the same time resolving to not repeat such behaviour in the next week, where both costs and benefits of the pub decision are weighted with similar discounting factors. Other cases of temptation can be captured in a similar way, with the decision-maker giving full weight to immediately enjoying a steak or cigarette while at the same time resolving to eat healthier or quit smoking in the far future, where costs and benefits of the decision are weighted with more or less similar discounting factors.

In what sense, then, are hyperbolic discounting theories good and successful models of dynamic inconsistency? In a general sense, hyperbolic discounting is parsimonious, predictively accurate and can be adopted in a wide variety of circumstances. Especially the theories of hyperbolic discounting for delay and quasi-hyperbolic discounting offer valuable extensions for models in consumer theory. Hence, in terms of prediction, hyperbolic discounting models of dynamic inconsistency have been successful.

In terms of explanation, hyperbolic discounting reduces dynamic inconsistency to a problem of attitudes to time. This is problematic for a number of reasons: first of all, as suggested in Chapter 4, time discounting is a complicated and not well-founded concept: often enough, the conceptual motivation for time discounting is unclear or ambiguous, i.e. it is unclear what exactly time discounting is supposed to capture. More specifically, while hyperbolic discounting can predict dynamic inconsistency, it does not offer much explanation in terms of how such dynamic inconsistency arises, i.e. how it is produced, what its mechanisms are, and what other factors it is influenced by.

This suggests that hyperbolic discounting approaches are not capable of explaining all aspects of more complex cases of dynamic inconsistency, such as temptation. Consider again the pub example: while dynamic inconsistency in this case can be predicted by hyperbolic discounting, it does not seem to fully capture the underlying mechanisms that lead to it. First of all, as demonstrated in Chapter 4, discounting needs to be underpinned by a substantial conceptual motivation to be well-founded, and to offer a satisfactory explanation in this context. Secondly, rather than because of short-sightedness that could be captured by hyperbolic discounting, the decision-maker could be motivated to go to the pub by various momentary desires, such as an urgent desire to drink, to converse with his friends, or to watch a rugby match. This line of critique is reinforced by Read (2006) who maintains that there are other sources of conflict than attitudes

to time.

Moreover, hyperbolic discounting is also ill-equipped to explain why agents differ in their dynamic inconsistency over different domains. Consider the case of a PhD student who drinks heavily, yet wants to quit drinking at the same time. The occupation of PhD student suggests that the decision-maker is capable of undertaking long-term projects, foregoing earnings in the short-run to have a more satisfying career in the long-run. At the same time, the PhD student cannot bring himself to quit drinking. While it is possible to postulate different discounting functions for different domains – an exponential one for his career and a hyperbolic one for drinking – it remains a limitation of the hyperbolic discounting model to not offer more specific resources to elucidate such differences in the dynamics of an agent's diverging interests.

Likewise, in terms of resolution of dynamic inconsistency, hyperbolic discounting is limited. As discussed in the above example, hyperbolic discounting can be used to predict dynamic inconsistency. However, such drastic changes in preference need to be supplied with a motivation in order to discuss how to resolve them. On this question, hyperbolic discounting theories do not lend themselves to give explanations as to why seemingly irrational behaviour as described above can be sufficiently well motivated. It is possible to view hyperbolic discounting as the expression of a taste for the immediate, yet, the above examples show how such a taste potentially undermines all other tastes an agent has. Offering little resources for understanding and explaining dynamic inconsistency, hyperbolic discounting is thus also not capable of formulating how agents might resolve dynamic inconsistency.

6.3.2 'Multi-Selves' Approaches

Complementing hyperbolic discounting theories, is a diverse literature which uses what Elster (1986) describes as the metaphor or idea of 'Faustian Selves' – the idea that there are two opposing groups of selves that make up the decision-maker, to explain and resolve dynamic inconsistency (as introduced in Chapter 2). The essence of those approaches is that they understand the decision-maker as populated by two (groups of) selves desiring different, mutually-exclusive outcomes and 'battling' for control over behaviour. To give an example, (Schelling, 1980, 58) says that

> 'people behave sometimes as if they had two selves, one who wants

> clean lungs and a long life and another who adores tobacco, or one who wants a lean body and another who wants dessert, or one who wants to improve himself by reading Adam Smith's theory of self-command and another who would rather watch an old movie on television.'

That is to say, such 'multi-selves' approaches seek to elucidate the inner conflicts of decision-makers that lead to dynamic inconsistency by identifying different (groups of) interests and motivations within them. Such approaches often provide different labels for the diverging interests in a decision-maker, including those of 'dual selves', 'hot and cold', 'planner-doer', 'strong and weak', 'far-sighted and short-sighted' and so on. The target phenomena of those approaches are first and foremost similar to those in the case of temptation. Indeed, this literature often refers to dynamic inconsistency as 'the problem of self-control'.

In the following, we review different specific proposals of 'multi-selves' accounts, starting with the 'picoeconomics' approach (Ainslie, 1975, 1992, 2001) that has foreshadowed the more formal models. The common starting point in those more formal models is that temptation and lack of self-control is modelled as a game between one 'planner'-self and many 'doer'-selves. Concerning such more formal models, we focus on the approach by Thaler and Shefrin (1981) followed by the approach of Fudenberg and Levine (2006).

Picoeconomics

Ainslie's 'picoeconomics' (Ainslie, 1992, 2001) describes intrapersonal bargaining processes with explicit reference to hyperbolic discounting.[3] His theory of 'picoeconomics' introduces the idea of micro-micro economics, i.e. the economics within one decision-maker. It can be understood as an investigation into the psychological mechanism that can underlie and produce hyperbolic discounting, supplementing the latter with a multi-self interpretation.

Ainslie (2005, 637) explicitly considers hyperbolic discounting (for delay) as a starting point for his theory:

> Hyperbolic discounting offers utility theory a rationale for why people should so frequently have impulses that contradict their own recognized best interests. These highly bowed curves shift the main prob-

[3]Indeed, Ainslie was one of the proponents of the delay theory of hyperbolic discounting which has been already used as a simple example of hyperbolic discounting as discussed in Chapter 4 and in Section 6.3.1 of this chapter.

> lem. We are no longer at a loss to explain choices that are short-sighted and temporary; now we have to account for how people learn the self-control that lets them adapt to a competitive world. How does an internal marketplace that disproportionately values immediate rewards grow into what can be mistaken for the long-range reward-maximizer of conventional utility theory?

The earlier analysis of hyperbolic discounting suggests that this statement is too quick: it is not clear in what sense hyperbolic discounting offers a rationale and an explanation for short-sightedness; all it offers is a modelling tool in which short-sighted behaviour can be predicted. To be sure, it is possible to supplement hyperbolic discounting with the right kind of interpretation, which gives a substantive meaning to the phenomenon of short-sightedness in agents (we will make a proposal for this in the next section). Hence, it seems that an account of self-control in the sense that Ainslie is aiming at will also have to include an explication of the processes that lead to present biases and temptation, before offering possible resolutions to those problems. Ainslie recognises this (implicitly), and develops a multi-selves account that explains hyperbolic discounting, as well as how it can be overcome:

> The orderly internal marketplace pictured by conventional utility theory becomes a bazaar of partially incompatible factions, where, in order to prevail, an option has not only to promise more than its competitors, but to act strategically to keep the competitors from later undermining it. The behaviors that are shaped by the competing rewards must deal not only with obstacles to getting their reward if chosen, but with the danger of being unchosen in favor of imminent alternatives. An agent [...] will be a succession of estimators whose conclusions differ; as time elapses these estimators shift their relationship with one another from cooperation on a common goal to competition for mutually exclusive goals. (Ainslie, 2005, 642)

That is, in Ainslie's view, a decision-maker becomes a collection of viewpoints that 'compete' for determining behaviour. This appears to support the intuition behind modelling the interaction between selves as dynamic games, which will be discussed in the next section. Furthermore, Ainslie maintains that the pattern

of the competition often results in behaviour that can be captured by hyperbolic discounting functions.

> Hyperbolic discount curves create a relationship of partial cooperation among your successive motivational states. Their individual interests in short range rewards, conflicting with their common interest in longer range rewards, create incentives much like those in the much studied bargaining game, repeated prisonerŠs dilemma. Choice of the better long range alternative at each point represents 'cooperation,' but this will look better than impulsive 'defection' only as long as you see it as necessary and sufficient to maintain your expectation that future selves will go on cooperating. (Ainslie, 2005, 642)

The picoeconomics, i.e. the bargaining processes between different motivational states, can hence be taken as elucidating how exactly individuals fail to control themselves and how they do so successfully. Bargaining processes and repeated games that Ainslie alludes to are not explicitly modelled by him. Rather, Ainslie (2001) provides different processes by which individuals can constrain themselves, including commitment strategies such as

- extrapsychic commitment, such as physically reducing future choice options (Ainslie, 2001, 74ff.),
- manipulation of attention, such as choosing to be ignorant (Ainslie, 2001, 76f.),
- preparation of emotion, such as cultivating the forestalling of affects (Ainslie, 2001, 77f.), and
- personal rules, such as making a resolution, and using willpower (Ainslie, 2001, 78ff.).

After outlining how such strategies can be successful in avoiding dynamic inconsistency (Ainslie, 2001, Chapter 6), he also investigates how such strategies can backfire, as rules can

- overshadow goods-in-themselves (Ainslie, 2001, 147f.),
- magnify lapses (Ainslie, 2001, 148f.),

- lead to misperception (Ainslie, 2001, 149ff.),
- lead to compulsive behaviour, (Ainslie, 2001, 155f.), and
- lead to an efficient will that can undermine appetite (Ainslie, 2001, 161ff.).

Along similar lines, though even less formally than Ainslie, Elster (2000) also considers strategies that are available to individuals to constrain their short-term action in order to achieve long-term goals.

In a general sense, Ainslie's theory describes possible resolutions of dynamic inconsistency and their respective problems from a psychological point of view. This suggests that perceiving of individuals as multiple selves can deepen our understanding of conflicts of motivations, especially in intertemporal settings.

We will now consider two models of the 'multi-selves' approach that aim to capture the processes described by Ainslie in a more formal way.

Dual Selves I: Competing Preferences

Thaler and Shefrin (1981, 394) model a decision-maker 'as having two sets of preferences that are in conflict at a single time'. One set of preferences is understood as the 'doer' and the other as the 'planner'. The planner is concerned with lifetime utility and the doer 'exists' only for one period and is completely myopic (this, in fact, implies a doer-self for each period in time). Furthermore, the doer is supposed to have direct control over the decision that is taken at the period at which he is active. At the same time, the planner has the possibility to influence the doer in certain ways. In this framework, present bias and lack of self-control in cases of temptation can be explained by an unconstrained doer-self maximising her utility in the given period. Introducing different selves which carry competing preferences hence formalises the above idea of conflicting motivations and explains dynamic inconsistency by giving up the idea of *synchronic* consistency.

The main objective of Thaler and Shefrin (1981) is to discuss in what way self-control can be achieved in the face of temptation, assuming the above setup. They introduce the idea of 'psychic technology' that the planner-self could use in order to constrain doer-selves and hence maximise lifetime utility. Two techniques are introduced:

> '(1) The doer can be given *discretion* in which case either his *preferences* must be modified or his *incentives* must be altered, or (2) the

> doer's set of choices may instead be limited by imposing *rules* that change the constraints the doer faces.' (Thaler and Shefrin, 1981, 395)

Concerning (1), the planner can exercise discretion by introducing self-binding mechanisms or plans (such as dieting programmes) or explicitly altering incentives (such as an alcoholic taking Antabuse which makes a person ill when drinking alcohol, or an academic agreeing to give a paper thus providing a proximate incentive to write it) (Thaler and Shefrin, 1981, 396f.). Such monitoring and persuasion is costly, and hence the planner can also (2) adopt rules that change the doer's constraints (such as spendthrifts imposing a ban on borrowing, dieters who never go to lavish dinner parties and gamblers who avoid Las Vegas) (Thaler and Shefrin, 1981, 397f.). Formally, such techniques are captured by a 'preference modification parameter' which changes the utility function of the doer such that they (are more likely to) choose according to the planner's utility maximisation. The lower the doer's consumption as a result of such modification is, the more modification according to one of the above resources is required.

Hence, in Thaler and Shefrin (1981), the synchronic consistency of the decision-maker is given up, allowing one to characterise present bias by the doer-self choosing the one-apple option and the planner-self endorsing the two-apples choices (and failing to influence the doer-self to comply with him). In the same vein, the temptation case can be elucidated, with the planner-self endorsing the overall assessment of the goodness of the three options and the doer-self undermining the assessment.

More recent models, such as Benabou and Pycia (2002), Fudenberg and Levine (2006), and Read (2006) build on this basic framework of two types of selves, modelling the processes by which decision-makers can exercise self-control in a much more detailed way. In particular, the modification techniques mentioned by Thaler and Shefrin (1981) are explicitly modelled as dynamic games between short-sighted and far-sighted selves in Benabou and Pycia (2002) and Fudenberg and Levine (2006).

Dual Selves II: Competing Time Horizons

Fudenberg and Levine (2006) model the interaction between a planner-self and doer-selves as a dynamic game. In their model, both types of selves have the *same* preferences, and only differ with regards to how they value the future. That is, in

contrast to Thaler and Shefrin (1981), their model retains synchronic consistency of the decision-maker and conceives of the diverging interests as one that is due to different time horizons of different selves.

Indeed, Fudenberg and Levine (2006, 1449) propose that 'many sorts of decision problems should be viewed as a game between a sequence of short-run selves and a long-run patient self.' The interaction between selves is modelled as a stage game. In the first stage, the long-run self chooses, at a cost for both the long-run self and the short-run self, an action that changes the utility function of the short-run self (similar to exercising discretion in Thaler and Shefrin (1981)). In the second stage, the short-run self, which is assumed to be completely myopic, makes a decision. This is repeated for each interaction between the long-run self and any short-term self. In this characterisation of the interaction between selves, the short-run self in each period has overlapping interests with the long-run self, as both share the same preferences over outcomes. However, the long-run self's utility is also determined by the outcomes of the stage-games with other short-run selves and will take an action in each of the first stages of the stage-games so as to maximise utility over all those interactions.

Consider again the two examples of present bias and temptation given earlier. Fudenberg and Levine (2006) retain synchronic consistency by postulating the same momentary preferences for both types of selves, yet endow the planner and doer with different time horizons. While this is a significant difference in modelling, the two examples are explained in very much the same way as with Thaler and Shefrin (1981). This suggests that both examples of dynamic inconsistency can be modelled, as well as explained, by those theories. Both types of models can capture different motivations than pure short-sightedness, by endowing different selves with competing evaluations or competing time horizons. For instance, concerning the temptation case, the models can capture the conflict in the decision-maker by conceiving of a doer-self that values getting drunk or watching rugby much more than avoiding the hangover, and a planner-self that values an avoided hangover. Furthermore, as suggested above, both theories render it possible to describe possible resolutions of dynamic inconsistency, by allowing the planner-self to adopt strategies to influence or constrain the doer-self. For instance, in the pub case, we can now conceive of the planner-self adopting such actions as to go out of the office with no money so as to not allow the doer-self to go into a pub. If such strategies of influencing (via constraints, incentives or pre-

commitment) are successful, dynamic inconsistency can be avoided. The much richer framework of the dual-self theories hence delivers additional insights into complex cases of dynamic inconsistency when compared to hyperbolic discounting theories.

Problems

We now turn to some problems that are implied by the above multi-selves frameworks. The above analysis suggests that the multi-selves models capture the idea advanced in Schelling (1980, 1984) that dynamic inconsistency can be explained by considering the interaction of selves. However, the way those models approach the task of modelling multi-selves or dual selves raises some problems.

Firstly, it can be asked what the above models add to the metaphor of the multiple-self. Loewenstein (1996, 288) maintains that we 'do not believe that there are little selves in people with independent motives, cognitive systems, and so on'. This, in turn, leads to the question of what exactly they add beyond the multiple-self metaphor. – We will show that by employing multiple-self models of personal identity over time, such concerns can be alleviated, as the latter can be motivated by substantial criteria of personal identity over time, as reviewed in Chapter 3.

Secondly, dual-self models claim to model the 'interaction' between selves. It is not immediately obvious what the term 'interaction' refers to here. Certainly, an individual can deliberate about courses of action from different perspectives, weighing the pros and cons according to the respective points of view. Yet, it is hard to see in what sense this amounts to an interaction between selves. A more convincing reading of the notion of interaction between selves is the idea that there are different points in time at which the different motivations can become actual. For instance, an individual that is planning to adhere to a more healthy diet can be depicted as having a planner-self that forms, endorses, and seeks to enforce actions that cohere with it. Moreover, the daily challenges of eating healthy can be depicted as temptations faced by the doers-selves that correspond to that point in time.

Thirdly, the approaches that use dynamic games to model interaction between selves are complex. As discussed in detail in Chapter 5, dynamic games bring with them an array of formal structure that needs to be interpreted. Even without endorsing any particular approach to game theory, all formal elements of the

extensive form need to be interpreted as depicting a game between opposing selves. Furthermore, when analysing such a dynamic game between selves from an epistemic perspective, the hypothetical reasoning of opponent-selves also needs to be captured by devices such as epistemic models, characterising beliefs and their revision, to finally motivate strategies in those games. Yet, it seems that in order to proceed in such a fashion, we would need to presume a higher degree of faction within a decision-maker to motivate the game-theoretic concepts (such as backward induction reasoning) than the degree of faction that we would like to explain. Yet, all this is needed if we are to draw on dynamic games as a modelling device. Maybe the conclusion to draw here is that we should not interpret the approach of modelling the interaction between selves as dynamic games too literally, adopting an 'as-if' interpretation of such models. However, while this will weaken the premise of endorsing extensive-form games between dual selves, it also weakens the extent to which such dual-self models can explain and resolve dynamic inconsistency. One of the very aims of the multi-selves approach was to explain (lack of) self-control connects, and the dynamic game models were supposed to provide a structure for this – one which is difficult to motivate in an explicit way.

Fourthly, due to their complexities, the dual-self models do not lend themselves to convenient and direct comparisons to normative decision theories. That is, the models do not reveal how much the different cases of dynamic inconsistency imply a departure from normative decision theory, which would give us a better understanding of the kinds of irrationality such cases bring with them.

More generally, the explanatory deficiencies of hyperbolic discounting theories and the complexities in the structure of the existing multi-selves approaches motivate the introduction of multiple-self models of personal identity over time to explain dynamic inconsistency in the next section.

6.4 Dynamic Inconsistency in Multiple-Self Models of Personal Identity over Time

In this section, we analyse problems of dynamic inconsistency with multiple-self models of personal identity over time. In particular, we introduce a simple multiple-self model which is consistent with hyperbolic discounting and an extended multiple-self model which is consistent with hyperbolic discounting and

in the spirit of the 'multi-self' models introduced in the preceding sections.

The main advantage of the multiple-self models of personal identity over time lies in the fact that their structure is simple, and that it is motivated by capturing of intertemporality for the deliberations of the decision-maker. This makes transparent what kind of assumptions in standard decision-theoretic representations need relaxation in order to capture dynamic inconsistency.

6.4.1 Present Bias in a Simple Multiple-Self Model

Recall the simple multiple-self model as introduced in Chapter 2, which postulates a set of temporal selves and characterises their degree of connectedness. Furthermore, as shown in Chapter 4, the model of a decision-maker as a collection of temporal selves can be used to motivate time discounting. Indeed, comparisons between the respective connectedness of temporal selves can be used to obtain values for time discounting functions. That is to say, time discounting can be motivated by the fact that the decision-maker considers changes in her future selves to influence her current evaluations.

To give a simple example, if all selves are different so that all temporal selves between each subsequent time points have a similar difference, a constant time discounting factor which gives the familiar exponential discounting function can be motivated. For this, one has to accept the assumption that under one of the connectedness interpretations, there is a constant degree according to which temporal selves change over time, i.e. S_2 differs from S_3 by the same degree as S_3 from S_4, which gives the same degree of connectedness $c_{2,3} = c_{3,4} = c_{t-1,t}$ and the connectedness between non-subsequent selves is obtained by combining the connectedness between all subsequent selves between them. Under these assumptions, a connectedness function can be constructed that reflects a uniformly behaved degree of connectedness. This type of exponential discounting due to connectedness was discussed in more detail detail in Sections 4.6.2.

As pointed out in Chapter 4, connectedness need not behave in such a uniform way. Diminishing psychological or empathy connectedness can behave in many different ways. In hyperbolic discounting, connectedness between subsequent selves in the near future diminishes faster than between those in the far future. More specifically, in order to endorse weightings given by hyperbolic discounting functions, connectedness between nearer selves needs to be perceived as relatively lower. Such an interpretation of hyperbolic discounting as reflecting one

of the two types of connectedness is an improvement over standard theories in the explanation of dynamic inconsistency. More formally, and in the terminology developed in Chapter 4, connectedness needs to motivate a time distance representation that coheres with a concave correspondence between time and clock-time. Indeed, rather than an ad hoc appeal to the fact that agents discount the future hyperbolically, the connectedness interpretations offer more detailed accounts as to why they might do so, namely because of diminishing psychological or empathy connectedness between their future selves.

Consider the example of present bias. In the psychological connectedness interpretation, the steeper discounting in the near future is due to a large perceived degree of taste change in the near future. It is intuitively plausible that an agent will know a lot more about her circumstances in the next few days and can easily depict many changes in tastes in this period of time, whereas her belief in the taste change in the far future will be much less differentiated. Hence, it could be the case that the agent does not discount the present at all (due to $c_{0,0} = 1$), yet the near future quite substantively (for instance, due to $c_{0,1} = .8$). This can lead to drastically devaluing an additional apple at t_1 (for instance because diet plans, activities, health, mood, etc. suggest many taste changes till then). However, the degree of taste change between two periods in the far future can be much less. For instance, an individual could foresee many taste changes in the far future, when compared to her tastes today. Accordingly, for the 999-th period she could adopt a discounting factor determined by $c_{0,999} = .1$. Now suppose the agent thinks about the likely taste changes on an even later date, yet she cannot perceive of a great difference between the two selves, so she might ascribe the same discounting factor to the the $1,000$-th period, due to $c_{0,1000} = .1$. Then, the additional apple in the far future is not so much devalued as the degree of taste change associated with the two selves in the far future is very similar. Hence, the choice of an apple today over two apples tomorrow, combined with the choice of two apples in $1,000$ days over one apple in 999 days can now be modelled. To continue with the example, on day 999, the agent evaluates her connectedness to her self at $1,000$ and now finds changes in taste as likely as between t_0 and t_1, hence ascribing $c_{0,1000} = .8$. As per the analysis given earlier, employing such discounting factors reveals dynamic inconsistency.

The example was analysed with psychological connectedness, but it is also possible to employ empathy connectedness, which offers broader possibilities for the

conceptual motivation of hyperbolic discounting. This suggests that the multiple-self model over time can motivate hyperbolic discounting, and its application in cases of dynamic inconsistency.

6.4.2 A Dual Multiple-Self Model

The model of temporal selves as introduced previously is limited in the sense that all conflict is, as in the hyperbolic discounting theories, reduced to a purely temporal one. However, there are also cases of dynamic inconsistency, such as the temptation cases, in which two points of view in a person are conflicting *at a time*, where one decision-maker has conflicting evaluations.

To model such cases, we introduce two (or more) 'rows' of temporal selves, which reflect the fact that there can be two competing points of view at a time (which are then each associated with a collection of temporal selves). That is, a decision-maker is modelled in a dual-self model, as introduced in Section 2.3.3 in Chapter 2. For simplicity, we will also refer to such a model as a 'two-row' model.

Firstly, there are two personalities, the planner and the doer, named P and D. Secondly, the two personalities have temporal selves associated with them. Hence, we have the following multiple-self model:

Time	$\mathbf{t_0}$	$\mathbf{t_1}$	...	$\mathbf{t_k}$
Planner	P_0	P_1	...	P_k
Doer	D_0	D_1	...	D_k

Table 6.1: Dual Selves in a Two-Row Model

The two rows of selves can now be interpreted analogously to the one-row model given earlier. That means that there are also two types of connectedness that can be used to characterise the temporal selves.

The two-row model will allow us to reconsider the recommendations of Thaler and Shefrin (1981) and Fudenberg and Levine (2006). While those models endorse similar structures of selves to the one above, here the additional tool of connectedness is introduced, which 'replaces' the modelling device of dynamic games.

We will comment in greater detail on the exact status of the two 'personalities' P and D and their associated selves when discussing the two cases just mentioned. However, as a general remark, recall that such multiple-self models

can be employed as devices that capture the deliberations of the decision-maker about the intertemporal aspects of prospects. That is to say, even the above two-row model can still be used in addition to a standard-decision theoretic representation that gives an evaluation, while the above model depicts a conflicting evaluation of the intertemporal aspects.

Temptation as a Connectedness Conflict

Indeed, in the Fudenberg and Levine (2006) model, the utility functions of the planner and the doer are assumed to be the same such that there is a standard decision-theoretic representation in the background. We show that by using the multiple-self model introduced above, the Fudenberg and Levine (2006) approach can be conceived as depicting a conflict in connectedness between selves. For this, we introduce an initial utility evaluation in which the planner and doer agree, and combine them with conflicting connectedness between the planner-selves and the doer-selves.

Consider the temptation example given earlier, and suppose that the utility of the two prospects is evaluated as follows.

	t_1	t_2
Pub	$u(\text{Pub}_1) = 10$	$u(\text{Pub}_2) = 2$
Home	$u(\text{Home}_1) = 2$	$u(\text{Home}_2) = 12$

Table 6.2: A Utility Evaluation

At t_0, the two prospects Pub and Home are evaluated, by ascribing utilities u to the individual consequences of each of the two prospects at t_1 and t_2, as depicted in the above table. Such an evaluation is consistent with the Fudenberg and Levine (2006) model in that it depicts a uniform evaluation of planner and doer. Note that the above evaluation assumes a separability of utility at different times. If the evaluation were to be concluded at this point without any further consideration, then the two prospects could be evaluated by their respective aggregate utilities, such as $u(\text{Pub}) = 12$ and $u(\text{Home}) = 14$. However, individuals often fail to maximise utility in those cases.

Following Fudenberg and Levine (2006), we consider that there are two selves, the planner and the doer, who have different time horizons. On the account developed here, this can be characterised as different degrees (and, possibly, kinds) of connectedness between the planner-selves and the doer-selves. Since the plan-

ner is concerned with lifetime utility, it is natural to endow him with a high degree of (psychological) connectedness c between all planner-selves. The doer, on the other hand, might be characterised as lacking empathy for different temporal selves, and hence can be endowed with low empathy connectedness c^*.

Analogous to the conclusions of the Fudenberg and Levine (2006) model, these different degrees of connectedness can then determine whether an individual chooses according to the planner or the doer. For this, we assume that it is possible to weight the goodness evaluation of prospects with the respective connectedness, as in the hyperbolic discounting approach. Since the utility evaluation above is endorsed by both the planner and the doer, all depends on the comparison of strength of connectedness c between the planner-selves and c^* between the doer-selves which will, in turn, have an impact on the final aggregation of the utilities for each prospect.

As a simple illustration, suppose the doer-selves are only perfectly connected between t_0 and t_1 and not at all connected between t_0 and t_2, such that $c^*_{0,1} = 1$ and $c^*_{0,2} = 0$. Further suppose that the planner-selves are high-connected such that $c_{0,1} = c_{0,2} = 1$, depicted in the below table.

	t_1	t_2
c	1	1
c^*	1	0

Table 6.3: Connectedness of Planner- and Doer-Selves

Weighting the above utility evaluation with connectedness, it makes no difference to the outcome whether we weight the utility evaluation with each of the two connectedness values and then aggregate the utilities for each prospect and comparing their values,[4] or whether we first consider aggregate connectedness and then weight the utility evaluation.[5] In both methods, the aggregate utility evaluation yields $u_A(\text{Pub}) = 11$ and $u_A(\text{Home}) = 8$ such that now going to the pub is preferred over going home, due to the extremely low connectedness between doer-selves. Yet, it is easy to see that for a better connected doer, the joint con-

[4]When adopting this method, we weight the utilities with c and c^* first. For the planner, the weighted utility evaluations are, as before, $u_P(\text{Pub}_1) = 10$, $u_P(\text{Pub}_2) = 2$, $u_P(\text{Home}_1) = 2$, $u_P(\text{Home}_2) = 12$. For the doer-self, the weighted utility evaluations are $u_D(\text{Pub}_1) = 10$, $u_D(\text{Pub}_2) = 0$, $u_D(\text{Home}_1) = 2$, $u_D(\text{Home}_2) = 0$. Aggregating those utilities yields $u_A(\text{Pub}) = 11$ and $u_A(\text{Home}) = 8$.

[5]When adopting this method, we consider the aggregate connectedness $c^A_{0,1} = 1$ and $c^A_{0,2} = .5$ and weight the utility evaluation such that $u_A(\text{Pub}_1) = 10$, $u_A(\text{Pub}_2) = 1$, $u_A(\text{Home}_1) = 2$, $u_A(\text{Home}_2) = 6$ which yields $u_A(\text{Pub}) = 11$ and $u_A(\text{Home}) = 8$.

nectedness could become higher such that eventually, going home becomes the recommended course of action.

In this context, consider that Xue (2008) develops hyperbolic discounting along similar lines, by depicting a decision-maker as two selves, one extremely myopic and one extremely far-sighted. More specifically, two (exponential) discounting functions are considered, one extremely myopic and the other extremely far-sighted (this roughly maps on to the above table with perfect connectedness c and imperfect connectedness c^*). Xue (2008) shows that aggregating those two discounting functions yields hyperbolic discounting. That is to say, all three of Fudenberg and Levine (2006), Xue (2008) and the above procedure allow us to analyse dynamic inconsistency in similar spirit.

In addition to what is offered by the models in Fudenberg and Levine (2006) and Xue (2008), connectedness gives a motivation for why the planner and the doer have a different outlook on the future. Indeed, the formulation adopted here also suggests that the interaction or bargaining models can be dispensed with when considering the different degrees and interpretations of connectedness, and to collapse the model back into one of hyperbolic discounting, where aggregate connectedness performs the role of a discounting function that is determined by c and c^*. Note though, how the connectedness interpretations permit us to make sense of successful self-control. For instance, stable tastes in a planner and some empathy in a doer can lead to weights that result in a higher aggregate utility for going home in the above example. Indeed, the above model permits us to formulate the exact conditions of connectedness for successful self-control, i.e. it permits us to investigate what combinations of planner- and doer-connectedness result in the aggregate utility for going home to be higher than in the pub case.

The above characterisation makes explicit a vital difference between the Fudenberg and Levine (2006) model and the Thaler and Shefrin (1981) model, which will be discussed shortly. In the former, the conflict between the selves is perceived as one of connectedness, and not as one of evaluation. That is, the goodness of each particular consequence at a time (such as a hangover, or a visit to the pub) is not disputed between the selves.

Temptation as a Utility Evaluation Conflict

Cases of temptation, including the above, can be interpreted as a more fundamental conflict in the evaluation of the actual consequences. It could be the case

that one self values being at home much more, and the other self perceives more vividly the hedonic pleasures of a pub visit. We now turn to an analysis of the above example that coheres with such a view.

In line with the Thaler and Shefrin (1981) model, the structure of the two-row model is now interpreted by considering two different utility functions for the planner and the doer. That is, there are two selves that have a genuine, synchronic, conflict in the evaluation of prospects, reflecting different desires. That is to say, the standard decision-theoretic representation is given up at this point in order to analyse deep conflicts in the decision-maker. More specifically, it is now assumed that the decision-maker has not one utility evaluation, but two utility evaluations that need to be reconciled.

We now consider a case in which the planner and the doer give the competing utility evaluations. Firstly, suppose the following utility evaluation of the planner:

	$\mathbf{t_1}$	$\mathbf{t_2}$
Pub	$u(\text{Pub}_1) = 10$	$u(\text{Pub}_2) = 2$
Home	$u(\text{Home}_1) = 2$	$u(\text{Home}_2) = 12$

Table 6.4: Utility Evaluation of the Planner-Self

As before, the two columns depict the distribution of utility over time. The table shows that the planner prefers to go home as $u(\text{Pub}) = 12$ is lower than $u(\text{Home}) = 14$. Now suppose the following utility evaluation of the doer:

	$\mathbf{t_1}$	$\mathbf{t_2}$
Pub	$u(\text{Pub}_1) = 12$	$u(\text{Pub}_2) = 2$
Home	$u(\text{Home}_1) = 2$	$u(\text{Home}_2) = 10$

Table 6.5: Utility Evaluation of the Doer-Self

The doer prefers to go to the pub, as $u(\text{Pub}) = 14$ is higher than $u(\text{Home}) = 12$. This decision-maker has hence a genuine conflict in the evaluation of the acts, such that his evaluations cannot be given by one overall utility evaluation. However, as per the example above, the decision-maker has developed two fully formed, yet conflicting points of view that can both be characterised as giving a utility evaluation.

Intuitively, we could also think of an individual with such a conflict as embodying different social roles which lead to different evaluations: for instance, evaluating the prospects from the perspective of being a father could yield a planner-type evaluation, and evaluating the prospect as someone who wants to

see an old friend in the pub might yield a doer-type evaluation. Both of these perspectives might be salient to the personality of the decision-maker and therefore capable of motivating the above evaluations.

Accepting that we are dealing with a decision-maker who has a deep evaluation conflict, we can attempt to resolve it. One possibility would simply be to aggregate the two evaluations. That is, the decision-maker is conflicted, but can still form two evaluations, and might consider to take them to account with equal weight. However, in the above example, such an aggregation does not yield a solution, as the aggregated utilities simply yield $u_A(\text{Pub}) = u_A(\text{Home}) = 13$ (when considering both views with equal weight).

However, we can still apply connectedness to the above evaluations. As per the dual-self model introduced earlier, we have again two degrees of connectedness, one for the planner, denoted c, and one for the doer, denoted c^*. As before, let the degree of connectedness between the selves at t_0 and the respective temporal selves at t_1 be perfect. That is, the connectedness between P_0 and P_1 is given by the unit weight, written $c_{0,1} = 1$, and likewise for the doer, $c^*_{0,1} = 1$.

Now suppose that the connectedness between the selves at t_0 and the respective temporal selves at t_2 are such that the doer is relatively lowly connected and the planner is relatively highly connected. Indeed, for simplicity, we assume perfect connectedness of the planner, such that $c_{0,2} = 1$ which means that we can still consider his initial utility evaluation as per above, even after weighting with connectedness. For the doer, we consider two cases. Firstly, let $c^*_{0,2} = .5$. Weighting his evaluation gives the following:

	$\mathbf{t_1}$	$\mathbf{t_2}$
Pub	$u(\text{Pub}_1) = 12$	$u(\text{Pub}_2) = 1$
Home	$u(\text{Home}_1) = 2$	$u(\text{Home}_2) = 5$

Table 6.6: Connectedness Weighting of the Doer's Evaluation I

That is, the consequences at t_2 are now devalued, such that the overall evaluation of the doer yields $u(\text{Pub}) = 13$ and $u(\text{Home}) = 7$. Secondly, consider the case of $c^*_{0,2} = 0$. Weighting his evaluation gives the following:

	$\mathbf{t_1}$	$\mathbf{t_2}$
Pub	$u(\text{Pub}_1) = 12$	$u(\text{Pub}_2) = 0$
Home	$u(\text{Home}_1) = 2$	$u(\text{Home}_2) = 0$

Table 6.7: Connectedness Weighting of the Doer's Evaluation II

Now, consequences at t_2 are completely devalued, such that the overall evaluation of the doer yields $u(\text{Pub}) = 12$ and $u(\text{Home}) = 2$.

We now have the planner's utility evaluation on the one hand, and two cases for the weighted utility evaluation of the doer. As before, we can now aggregate the different evaluations. Yet, different methods of aggregating the weighted evaluations of planner and doer can lead to different results. We will discuss which method of aggregation to select in such situations in the next section.

Aggregation Conflicts

Here we consider two possible ways of aggregating the weighted evaluations of the two selves. We first discuss what result those methods yield, before considering arguments for and against each of these methods.

Aggregation over acts. Firstly, we could simply aggregate over *acts*, i.e. we add the respective evaluations of going to the pub and going home and divide by the number of selves (in this case two). That is to say, if a decision-maker has a deep conflict and has two opposing viewpoints, then she could first weight the evaluations of those viewpoints with their associated connectedness, then aggregate over times, and finally aggregate the evaluations over the different acts. This method, applied for each of the different connectedness cases considered here, yields the following table.

	Initial	$c^*_{0,2} = .5$	$c^*_{0,2} = 0$
$u_A(\text{Pub})$	13	12.5	12
$u_A(\text{Home})$	13	10.5	8

Table 6.8: Aggregation over Acts

For the initial utility evaluations, we get the aggregate utilities of $u_A(\text{Pub}) = u_A(\text{Home}) = 13$, as mentioned before. This is depicted in the first column of the above table. For the case of moderate doer-connectedness ($c^*_{0,2} = .5$), we get $u_A(\text{Pub}) = 12.5$ and $u_A(\text{Home}) = 10.5$, as depicted in column two. For the case of zero doer-connectedness, we get $u_A(\text{Pub}) = 12$ and $u_A(\text{Home}) = 8$, shown in column three. Hence, on this method of aggregation, low doer-connectedness tilts the overall utility evaluation in favour of going to the pub.

Aggregation over selves. Secondly, we could aggregate over *selves*. That is, we are comparing the four possible paths, considering the utilities that each self associates with the two different acts. Then, for each self, the connectedness

weight is applied and the evaluation of each act is aggregated over times. Finally, the act that yields the highest utility is selected for each self, and compared in the below table for the different cases of connectedness considered hre.

	Initial	$c^*_{0,2} = .5$	$c^*_{0,2} = 0$
Doer	14	13	12
Planner	14	14	14

Table 6.9: Aggregation over Selves

We can now check whether there is a dominating act-self combination. For the initial evaluation, we have the planner valuing going home with 14 utils, and likewise the doer valuing going to the pub with 14 utils. In each of the weighted cases, however, the evaluations of the doer are lowered, such that the planner's evaluation of going home is the best overall act under this method.

How are we to decide between these methods of aggregation? The first method assumes that we can simply add the utilities of both evaluations. This is plausible, because the initial evaluation conflict suggests that we have really do have two equally viable points of view in the decision-maker. Indeed, this conflict has induced two rows of selves and their utility functions, which in turn makes it also plausible to consider connectedness separately. After weighting the evaluations with their respective connectedness, it is more plausible to assume intrapersonal comparability of utility (as differences in the future stability of the evaluations are now taken into account), and aggregation yields an overall evaluation.

However, in the above example, this means that because of low connectedness of the doer, the option of going home is valued less, and hence the pub is recommended. While this might ring true descriptively, it does not seem to be a particularly good reason not to go home: we would expect that the low connectedness of the doer results in his perspective counting less than the high-connected planner. – Yet, the objection against the low connectedness of the doer having an influence on the decision is flawed. It seems that it rests on the intuition that the doer's perspective is not on a par with the planner's perspective. If that is the case, then the model with two conflicting utility functions is misapplied and the case should simply be modelled as a conflict in connectedness.

Temptation as a Test of Character

This leaves the question of whether the second method of aggregation should be preferred. The first method of aggregation takes the connectedness-weighted utilities at face value and aggregates over acts. The second method of aggregation also appears to take the utilities at face value, indeed, it seems to take the utilities almost literally. Instead of asking which act is the best one, it is asking which of the selves has the best evaluation of an act.

Note how the second method of aggregation implies a slight, yet important, change in the question. The above exercise has started as an analysis of conflicting evaluations over acts, asking how to reconcile them by aggregating the evaluations after weighting them with connectedness. That is, the initial question has considered the best act (or prospect) in terms of its consequences, and the first method of aggregation has used connectedness to identify it by weighting the compating evaluations of the available acts. Yet, the second method of aggregation analyses competing *evaluators*. That is, the question has shifted from considering the 'best act' to considering the 'best self', i.e. the one which yields the best evaluations. In the second method, we are hence asking which perspective is superior – and if one of the perspectives is associated with lower connectedness, it will count less in the aggregation.

In terms of the temptation case, this change in question is quite telling: intuitively, if a decision-maker is able to conceive of a temptation he faces as a test of character, then successful self-control can be described by the third method of aggregation. Cases of dynamic inconsistency, i.e. lack of self-control in the face of temptation, can be captured by low connectedness of the doer, combined with the first method of aggregation. That is, if a decision-maker deliberates about the problem and tries to decide between going home or going to the pub in this way, he will go to the pub. However, when a decision-maker deliberates for a longer time about this problem – which also makes postulating two selves with different utility functions more plausible – then the deliberations might not stay exclusively with considering the goodness of the acts and the future stability of the desires.

Indeed, it is natural that in such cases, the decision-maker might ask himself what kind of character he wants to be: someone who enjoys hedonic pleasures or someone who is prudent? If a case of temptation leads a decision-maker to entertain such more fundamental questions, then it is natural that the second method

of aggregation becomes more plausible, as it answers the question which character or standpoint will maximise his overall utility, and the low connectedness of one of those standpoints has the impact of lowering the influence of that particular standpoint. Especially if one of the selves has significantly higher connectedness, this could be taken as indicator of the stability of that viewpoint when it comes to implementing the recommendation it gives. That is to say, successful self-control in the face of temptation can be explained by the decision-maker perceiving of such situations as a test of character.

This concludes the analysis of dynamic inconsistency and intrapersonal conflicts by multiple-self models of personal identity over time. In a general sense, the above discussion suggests that the two-row model can describe present bias, temptation cases and deeper kinds of intrapersonal conflicts either in terms of connectedness conflicts or in terms of evaluation conflicts. In the connectedness conflict, sufficiently low connectedness of the doer can lead to present bias and failure of self-control, and a sufficiently high degree of connectedness of the planner can yield successful self-control. Evaluation conflicts allow us to model deliberation of a decision-maker who has a synchronic conflict between different standpoints of preferences: namely, we can consider introspection about how likely it is that tastes that are associated with a particular point of view are going to change. The three methods of aggregation capture the differences between failure and successful self-control in the face of temptation. Moreover, the third method of aggregation also allows us to characterise deeper kinds of intrapersonal conflict that lead the decision-maker to deliberate about character planning.

6.4.3 Theories of Dynamic Inconsistency Revisited

Here we reconsider the problems and deficiencies in the hyperbolic discounting and multi-selves models in the light of the multiple-self models of personal identity over time offered in this section.

Firstly, we questioned what the models add to the metaphor of the multiple-self. In the multiple-self model of personal identity over time, the metaphor is cashed out in terms of a decision-maker's stability over time, characterising the degree of connectedness between selves. It was shown in Chapter 3 that the structure of the multiple-self models coheres with theories of personal identity over time, such that they can be used to motivate and constrain the models, for instance by introducing the interpretations of psychological and empathy connec-

tedness. This does not mean that the models have a fundamental metaphysical character. Rather, it suggests that the objects and concepts it poses have a substantial interpretation, which can be a very reductive one, such as when considering selves as sets of preferences and connectedness as a rough measure of diachronic similarity of preferences.

Secondly, the multi-self models in the literature are complex, postulate additional structure in the model of the decision-maker. This has also made it difficult to understand to what extent they require us to depart from standard decision-theoretic accounts. Introducing the notion of temporal selves and their connectedness makes transparent how those models relates back to standard decision-theoretic representations of decision-makers and in what sense those models can be endorsed as picturing rational agency.

For instance, it is a relatively small departure from standard decision theory to consider one row of temporal selves, measuring the similarity of their tastes to determine connectedness. Each further step that can be taken in the model, such as using connectedness as discounting factors, and introducing further rows, and non-reductive interpretations of selves makes clear to what extent standard decision theory is enriched, and departed from. Connectedness conflicts in a two-row model are a more significant departure: the way evaluation conflicts are modelled and the aggregation conflicts they raise demonstrate that conflicts which induce different points of view that are modelled as competing utility functions are a rather significant departure from standard decision theories. Nevertheless, they still offer an analysis of such cases of deep intrapersonal conflicts, which permits us to understand in what way exactly standard assumptions of decision theory need to be relaxed.

Finally, while the above multiple-self model of personal identity over time is not dynamic in a direct sense, note how it is capable of modelling the differences between success and failure in self-control. Moreover, the dynamic aspect of accounts such as Ainslie (1992, 2001) and Fudenberg and Levine (2006) is also partly due to modelling hypothetical repetitions, which is also possible with the above model. For instance, by considering repeated interactions between planners and doers, more introspective evidence is amassed about the connectedness associated with each perspective. Successfully executed plans, for instance, can be used to update the connectedness of the planner. Likewise, an individual could learn over time what kinds of hedonic pleasures are stable ones, boosting

the connectedness of the doer. This suggests that the above model can also be considered in a dynamic sense.

6.5 Conclusions

We have analysed a particular type of preference change in intertemporal decisions, namely that of dynamic inconsistency. Received accounts in behavioural economics that attempt to predict, explain, and resolve dynamic inconsistency have been critically scrutinised. The more general model of connectedness in the multiple-self has been used to re-describe some of the features of the aforementioned accounts and their recommendations, which has led to some improvements: the multiple-self model makes it possible to better motivate the additional structure, relating it back to standard decision theories.

In a general sense, as mentioned in the introduction to this chapter, the multiple-self model allows us to more carefully consider to what extent (hyperbolic) discounting approaches can be employed to model dynamic inconsistency. It has been shown that if such models are developedto reflect connectedness in the multiple-self, they can even be used to analyse conflicts in evaluation. That is to say, discounting functions which reflect a coarse-grained evaluation of distance in time can, when interpreted with connectedness, and when applied to rows of conflicting selves, even be used to elucidate aspects of complicated cases of preference conflicts. To be sure, not all preference conflicts have an explicit intertemporal dimension, and even if they have, there is no guarantee that connectedness will conclusively analyse or even resolve the conflict. Yet, many preference conflicts that are cases of dynamic inconsistency can be analysed fruitfully with connectedness.

Conclusions

Chapter 7

Conclusions

7.1 Time in Decisions and Games

This thesis has been concerned with the analysis of intertemporality in decisions and games. We have identified three questions about intertemporality in decisions and games, namely the problems of (i) the correct evaluation of temporal distance by time discounting functions in Chapter 4, (ii) belief revision about other decision-makers in interaction over time by backward induction reasoning in Chapter 5, and (iii) preference change by accounts of dynamic inconsistency in Chapter 6.

Concerning the evaluation of temporal distance by time discounting function, we have identified two goals that theories of time discounting may have: one, postulating a correct time discounting function, and two, offering an accurate underlying conceptual motivation. Since both of those concerns can be understood either descriptively or normatively, this creates four problems of time discounting. Reviewing existing theories of time discounting and their representation frameworks, we have seen that it is difficult to state in what sense they address the four goals of time discounting theories, as their frameworks of representation entangle matters of goodness evaluation on the one hand and time feature evaluation on the other hand. We proceeded by presenting a general representation framework for time discounting which initially separates goodness and time feature evaluations. Adopting a measurement-theoretic framework, we outlined the requirements that time discounting functions have to fulfill in order to be well-founded in a numerical representation of the qualitative notion of time distance. Furthermore, to state discounting functions as time-indexed weights, such a representation has

to correspond to an externally given time-index. On the basis of this general framework, existing theories of time discounting have been critically scrutinised, making their underlying assumptions explicit and raising questions about their conceptual motivations. In a further step, we have shown that Parfit's dictum of time discounting because of a weak connectedness to future selves in a decision-maker can motivate the general representation, which has enabled us to address objections against his view. Furthermore, the general framework allows to derive exponential discounting from a notion of preference change. More generally, the requirements for time discounting theories developed here demonstrate that time discounting factors are restricted in the kinds of conceptions they can express.

Concerning interaction between decision-makers over time, we have analysed the reasoning method of backward induction that is standard in dynamic games. We analysed the sequential structure of dynamic games with perfect information. A three-stage account was proposed, that specifies set-up, reasoning and play stages of dynamic games. Accordingly, we have defined a player as a set of agents corresponding to these three stages. Moreover, the notion of agent connectedness was introduced which measures the extent to which agents' choices are sequentially stable. In a next step, a type-based epistemic model was augmented with agent connectedness and used to provide sufficient conditions for backward induction. Besides, an existence result was given to ensure that our conditions are indeed possible. Our epistemic foundation for backward induction makes explicit that the epistemic independence assumption involved in backward induction reasoning is stronger than usually presumed. Furthermore, in the three-stage account, players can explicitly be understood as multiple-selves, which makes available connectedness accounts motivated by theories of personal identity over time. Thus describing dynamic games allows a more fully understanding of strategic interaction over time and gives a more realistic representation of players for game-theoretic applications in economics and social sciences.

Concerning preference change over time, we have analysed theories of dynamic inconsistency. Two families of approaches can be identified: those that use hyperbolic discounting functions to describe dynamically inconsistent decision-makers as myopic, and those that postulate multi-selves models that capture different motivations and time horizons which can lead a decision-maker to (fail to) control himself in the face of temptation. We have analysed those accounts and have pointed out that hyperbolic discounting functions have explanatory deficiencies

as they reduce it to a specific evaluation of temporal distance. Concerning multi-selves models, we have shown that their structure is complex, as they employ dynamic games (analysed in Chapter 5) to model the interaction between selves. In order to achieve a simpler characterisation of dynamic inconsistency, we have reconsidered both hyperbolic discounting and multi-selves models in the more general multiple-self model of personal identity over time. A simple specification of it can motivate hyperbolic discounting, and an extended version of it has been used to reformulate the multi-selves models by using less formal structure that can be better motivated. Moreover, the latter allows to distinguish between conflicts in connectedness and conflicts in evaluation, with the latter also being able to enhance our understanding of intrapersonal conflicts more generally.

In general, analysing those three aspects of intertemporality in decisions and games has shown that while their interrelations are intricate and plentiful, it is important to analyse those different concerns separately, at least in an initial step. In particular, ad hoc appeals to time discounting are problematic in other areas of analysis, such as in dynamic games or in the analysis of preference change. This is due to the fact that many regularity assumptions are required in order to construct well-founded time discounting functions, as demonstrated by the general representation framework of time discounting in Chapter 4. Furthermore, standard frameworks of describing interaction over time contain many implicit stability assumptions, such as those about the stability of preferences and strategies in extensive-form games, as analysed in Chapter 5. Making such stability assumptions explicit and presenting models that allow us to relax those, and consider specific explanations for why stability breaks down (or how it can hold) improves our understanding of interaction over time. In a similar vein, the analysis of preference change in Chapter 6 has shown how simple and dual multiple-self models can capture explain and resolve dynamic inconsistency, making transparent the assumptions that we need to relax in normative decision theories in order to model dynamic inconsistency.

7.2 The Multiple-Self in Decisions and Games

The three accounts of intertemporality in decisions and games presented in this thesis have been facilitated by introducing multiple-self models of personal identity over time. While most of the formal results and main arguments in the

aforementioned accounts do not *depend* on endorsing such devices as extensions of standard decision theories, we found that the analysis of each of the problems has been enriched by their explicit consideration.

The multiple-self models of personal identity over time have two main elements, as introduced in Chapter 2. In a first step, we view the decision-maker as a collection of temporal selves. In a second step, the degree of connectedness between the temporal selves gives a characterisation of the decision-maker's stability over time. We have shown in Chapter 3 how such multiple-self models structurally cohere with theories of personal identity over time, which are concerned with how persons persist and change over time. By characterising the connectedness between selves, a multiple-self model can be interpreted and further constrained by specific accounts of theories of personal identity over time.

Through applying the multiple-self models of personal identity over time in the three accounts of intertemporality in decisions and games, many specific interpretations have become available. Firstly, we have shown in what sense time discounting can be motivated by diminishing connectedness between selves. Secondly, we have shown how connectedness between different selves can help to locally relax stability assumptions in dynamic games, and we have motivated new sufficient conditions for backward induction with this notion. Thirdly, re-describing theories of dynamic inconsistency with multiple-self models better motivates their often complicated structure and renders clear in what sense they require us to depart from normative decision theory.

Intertemporality in decisions and games, it seems, calls for what Sen (1977) has labelled as a need for 'more structure' in decision theory. While this assertion was made in the context of pointing out deficiencies in some of the descriptive and ethical implications of standard normative decision theories, it is a fitting slogan for the multiple-self models developed in this thesis. Two initial observations have motivated their development. Firstly, standard decision theories are not geared towards explicitly taking into account the temporal dimension of decisions in greater detail, as shown in Chapter 2. Secondly, the extensive literature on topics such as time discounting, dynamic games and dynamic inconsistency suggests that time is a key factor in decisions and games that is often subjected to separate analysis. This raised the question of how such analyses relate to the foundations of decision and game theory. It is hoped that the multiple-self framework developed in this thesis contributes to the investigation of this ques-

tion. Indeed, the multiple-self model has been developed so as to be able to track what kinds of assumptions in addition to standard decision-theoretic frameworks need to be given up in order to analyse intertemporal decisions and games with regards to the three problems identified in the literature.

7.3 Future Work

The simple, and yet general, character of the multiple-self models of personal over time developed in this thesis offers various routes for further research.

One route is to generalise the idea of connectedness between selves to one of connectedness between individuals to study distance between individuals in social decisions. In particular, amending utilitarian frameworks with a notion of connectedness could explain the effect of spatial distance in failures to maximise aggregate utility. Furthermore, by modelling group decision-making in the multiple-self, different formal specifications of connectedness can be related to existing frameworks of deliberation and aggregation.

Offering further accounts of problems of intertemporality is another possible route, such as (i) in-depth characterisations of specific time discounting functions, and extending the general representation framework for time discounting to intergenerational decisions in which population changes suggest shifting domains, (ii) characterisations of backward induction in an imperfect information framework, and characterisations of forward induction, and (iii) exploring the links between connectedness and recent models of preference change in philosophical decision theory, as well as further investigating how connectedness can be used to elucidate the differences between descriptive and normative decision theory.

In general, I hope to have shown that connectedness models of intertemporality can improve our understanding of the role of time in rational decision-making, in virtue of elucidating the accounts of time discounting, backward induction, and dynamic inconsistency which have been proposed as answers to the three questions of time in decisions and games considered in this thesis.

Bibliography

Adams, E. W. (1966). On the nature and purpose of measurement. *Synthese*, **16**, 125–169.

Ainslie, G. (1975). Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin*, **82**(4), 463–96.

Ainslie, G. (1992). *Picoeconomics: the Interaction of Successive Motivational States within the Individual*. Cambridge University Press.

Ainslie, G. (2001). *Breakdown of Will*. Cambridge University Press.

Ainslie, G. (2005). Precis of breakdown of will. *Behavioral and Brain Sciences*, **28**, 635–73.

Akerlof, G. A. and Kranton, R. E. (2000). Economics and identity. *Quarterly Journal of Economics*, **115**(3), 715–753.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine. *Econometrica*, **21**, 503–546.

Angeletos, G.-M., Laibson, D., Repetto, A., Tobacman, J., and Weinberg, S. (2001). The hyperbolic consumption model: Calibration, simulation, and empirical evaluation. *Journal of Economic Perspectives*, **15**, 47–68.

Aumann, R. J. (1976). Agreeing to disagree. *The Annals of Statistics*, **4**, 1236–1239.

Aumann, R. J. (1987). Correlated equilibrium as an expression of bayesian rationality. *Econometrica*, **55**, 1–18.

Aumann, R. J. (1995). Backward induction and common knowledge of rationality. *Games and Economic Behavior*, **8**, 6–19.

Bach, C. W. and Heilmann, C. (2009). Agent connectedness and backward induction. *LSE Choice Group Working Paper Series*, **5**(3).

Barro, R. J. (1999). Ramsey meets Laibson in the neoclassical growth model. *Quarterly Journal of Economics*, **114**(4), 1125–1152.

Battigalli, P. and Bonnano, G. (1999). Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics*, **53**, 149–225.

Battigalli, P. and Siniscalchi, M. (2002). Strong belief and forward induction reasoning. *Journal of Economic Theory*, **106**, 356–391.

Belzer, M. (2005). Self-conception and personal identity: Revisiting Parfit and Lewis with an eye on the grip of the unitary reaction. *Social Philosophy and Policy*, **22**(2), 126–164.

Benabou, R. and Pycia, M. (2002). Dynamic inconsistency and self-control: a planner-doer interpretation. *Economics Letters*, **77**(3), 419–424.

Binmore, K. (1987). Modeling rational players I. *Economics and Philosophy*, **3**, 179–214.

Board, O. (2002). Knoweldge, beliefs and game-theoretic solution concepts. *Oxford Review of Economic Policy*, **55**, 433–445.

Boumans, M. (2007). *Measurement in Economics: A Handbook*. AP Elsevier.

Bradley, R. (2004). Ramsey's representation theorem. *Dialectica*, **4**, 484–497.

Bradley, R. (2007a). The kinematics of belief and desire. *Synthese*, **56**(3), 513–535.

Bradley, R. (2007b). A unified Bayesian decision theory. *Theory and Decision*, **63**, 233–263.

Bradley, R. (2009a). Becker's thesis and three models of preference change. *Politics, Philosophy and Economics*, **8**(2), 223–242.

Bradley, R. (2009b). Preference kinematics. In T. Grüne-Yanoff and S. O. Hansson, editors, *Preference Change*, pages 221–242. Theory and Decision Library A, Springer.

Bradley, R. (2009c). Revising incomplete attitudes. *LSE, mimeo.*

Brandenburger, A. (1992). Knoweldge and equilibrium in games. *Journal of Economic Perspectives*, **6**, 83–101.

Brandenburger, A. (2007). The power of paradox: some recent developments in interactive epistemology. *International Journal of Game Theory*, **35**(4), 465–492.

Brandenburger, A., Friedenberg, A., and Keisler, H. J. (2008). Admissibility in games. *Econometrica*, **76**, 307–352.

Bratman, M. (1996). Planning and temptation. In L. May, M. Friedman, and A. Clark, editors, *Mind and Morals*, chapter Planning and Temptation. MIT Press.

Broome, J. (1991). *Weighing Goods*. Basil Blackwell.

Broome, J. (1999). *Ethics out of Economics*. Cambridge University Press.

Cassam, Q. (1999). *Self and World*. Oxford University Press.

Chisholm, R. (1976). *Person and Object*. La Salle: Open Court.

Connor, K., de Dreu, C., Scroth, H., Barry, B., Lituchy, T., and Bazerman, M. H. (2002). What we want to do versus what we think we should do: An empirical investigation of intrapersonal conflict. *Journal of Behavioral Decision Making*, **15**(5), 403–418.

Dancy, J. (1997). *Reading Parfit*. Oxford: Blackwell Publishers.

Davidson, D. (1980). *Essays on Actions and Events*. Cambridge: Cambridge University Press.

de Bruin, B. (2009). Overmathematisation in game theory: Pitting the Nash equilibrium refinement programme against the epistemic programme. *Studies In History and Philosophy of Science Part A*, **40**(3), 290–300.

DeGrazia, D. (2005). *Human Identity and Bioethics*. Cambridge University Press.

Descartes, R. (1637). *Discourse on Method*. Available online: http://www.gutenberg.org/files/59/59-h/59-h.htm.

Dietrich, F. and List, C. (2009). A model of non-informational preference change. *LSE Choice Group Working Paper Series*, **5**(1).

Elster, J. (1986). *The Multiple Self*. Cambridge University Press.

Elster, J. (2000). *Ulysses Unbound*. Cambridge University Press.

Evnine, S. J. (2008). *Epistemic Dimensions of Personhood*. Oxford University Press.

Fishburn, P. C. and Rubinstein, A. (1982). Time preference. *International Economic Review*, **23**(3), 677–94.

Frederick, S. (1999). *Discounting, Time Preference, and Identity*. Ph.D. thesis, Carnegie Mellon University.

Frederick, S., Loewenstein, G., and O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, **40**(2), 351–401.

Fudenberg, D. and Levine, D. K. (2006). A dual-self model of impulse control. *American Economic Review*, **96**(5), 1449–1476.

Gallois, A. (2008). Identity over time. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, winter 2008 edition.

Gendler, T. S. (2000). *Thought Experiment: On the Powers and Limits of Imaginary Cases*. New York: Garland Press.

Gollier, C. (2002). Discounting an uncertain future. *Journal of Public Economics*, **85**, 149–166.

Halevy, Y. (2008). Strotz meets Allais: Diminishing impatience and the certainty effect. *American Economic Review*, **98**(3), 1145–1162.

Halpern, J. (2001). Substantive rationality and backward induction. *Games and Economic Behavior*, **37**, 425–435.

Hammond, P. J. (1976). Changing tastes and coherent dynamic choice. *The Review of Economic Studies*, **43**(1), 159–173.

Harsanyi, J. C. (1967). Games of incomplete information played by "Bayesian players". Part I, II, III. *Management Science*, **14**, 159–182, 320–334, 486–502.

Herrnstein, R. (1981). Self-control as response strength. In C. M. B. et al, editor, *Quantification of Steady-State Operant Behavior*, pages 3–20. Elsevier.

Howson, C. (1997). Bayesian rules of updating. *Erkenntnis*, **45**, 195–208.

Hume, D. (1739). *A Treatise of Human Nature*. Clarendon, Oxford.

Jeffrey, R. C. (1983). *The Logic of Decision*. University of Chicago Press.

Joyce, J. (1999). *The Foundations of Causal Decision Theory*. Cambridge University Press.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, **47**, 263–291.

Kolak, D. and Martin, R. (1991). *Self and Identity: Contemporary Philosophical Issues*. New York: Macmillan.

Koopmans, T. (1960). Stationary ordinal utility and impatience. *Econometrica*, **28**(2), 287–309.

Korsgaard, C. M. (1989). Personal identity and the unity of agency. *Philosophy & Public Affairs*, **18**(2), 101–32.

Krantz, D. H., Luce, R. D., Tversky, A., and Suppes, P. (1971). *Foundations of Measurement Volume I: Additive and Polynomial Representations*. Mineola: Dover Publications.

Kuhn, H. W. (1953). Extensive games and the problem of information. *Annals of Mathematics Studies*, **28**, 193–216.

Laibson, D. (1986). Self-control and saving for retirement. *Brookings Papers on Economic Activity*, **1**, 91–196.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics*, **112**(2), 443–477.

Lancaster, K. (1963). An axiomatic theory of consumer time preference. *International Economic Review*, **4**(2), 221–231.

Levi, I. (1986). *Hard Choices: Decision Making under Unresolved Conflict*. Cambridge University Press.

Lewis, D. (1983). Survival and identity. In R. Martin and J. Barresi, editors, *Personal Identity*, pages 114–167. Blackwell Publishing, 2003.

Locke, J. (1694). *An Essay Concerning Human Understanding*. Book II, ch. XXVII.

Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, **65**(3), 272–292.

Loewenstein, G. and Elster, J. (1992). *Choice over Time*. Russell Sage.

Loewenstein, G. and Prelec, D. (1992). Anomalies in intertemporal choice: Evidence and an interpretation. *Quarterly Journal of Economics*, **107**(2), 573–597.

Loewenstein, G. and Read, D. (2003). *Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice*. Russell Sage.

Loomes, G. and Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal*, **92**(368), 805–824.

Lorenz, H. (2009). Ancient theories of soul. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, summer 2009 edition.

Lowe, J. E. (2001). *The Possibility of Metaphysics - Substance, Identity, and Time*. Oxford: Oxford University Press.

Luce, R. D., Krantz, D. H., Tversky, A., and Suppes, P. (1971). *Foundations of Measurement Volume III: Representation, Axiomatization, and Invariance.* Mineola: Dover Publications.

MacIntyre, A. (1984). *After Virtue.* Notre Dame: University of Notre Dame Press.

MacIntyre, A. (1989). The virtues, the unity of a human life and the concept of a tradition. In S. Hauerwas and L. G. Jones, editors, *Why Narrative?* Grand Rapids, MI: W.B. Eerdmans.

Manzini, P. and Mariotti, M. (2007). Choice over time. *Institute for the Study of Labor (IZA) Discussion Papers No. 2993.*

Martin, R. and Barresi, J. (2003). *Personal Identity.* Blackwell Publishing.

Mas-Colell, A., Whinston, M. D., and Green, J. R. (1995). *Microeconomic Theory.* Oxford University Press, USA.

Mazur, J. (1987). An adjustment procedure for studying delayed reinforcement. In C. et al, editor, *The Effect of Delay and Intervening Events on Reinforcement Value*, pages 55–73. Hillsdale: Erlbaum.

McClennen, E. F. (1990). *Rationality and Dynamic Choice: Foundational Explorations.* Cambridge University Press, Cambridge.

Nagel, T. (1971). Brain bisection and the unity of consciousness. In J. Perry, editor, *Personal Identity*, pages 227–245. University of California Press, Berkeley, 1975.

Nash, J. F. (1951). Noncooperative games. *Annals of Mathematics*, **54**, 289–295.

Noonan, H. W. (1989). *Personal Identity.* Routledge.

Noonan, H. W. (2008). Identity. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Stanford University, fall 2008 edition.

Nozick, R. (1981). *Philosophical Explanations.* Oxford: Clarendon Press.

O'Brien, L. (2007). *Self-Knowing Agents.* Oxford University Press.

Ok, E. A. and Masatlioglu, Y. (2007). A theory of (relative) discounting. *Journal of Economic Theory*, **137**, 214–45.

Olson, E. (1997). *The Human Animal: Personal Identity Without Psychology.* Oxford: Oxford University Press.

Olson, E. (2003). An argument for animalism. In R. Martin and J. Barresi, editors, *Personal Identity.* Blackwell Publishing.

Olson, E. T. (2008). Personal identity. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, winter 2008 edition.

Parfit, D. (1971). Personal identity. *The Philosophical Review*, **80**(1), 3–27.

Parfit, D. (1984). *Reasons and Persons*. Clarendon.

Peleg, B. and Yaari, M. E. (1973). On the existence of a consistent course of action when tastes are changing. *The Review of Economic Studies*, **40**(3), 391–401.

Perea, A. (2007). Epistemic conditions for backward induction: An overview. In *Interactive Logic Proceedings of the 7th Augustus de Morgan Workshop*, pages 195–193. Texts in Logic and Games 1, Amsterdam University Press.

Perea, A. (2008). Minimal belief revision leads to backward induction. *Mathematical Social Sciences*, **56**, 1–26.

Perea, A. (2011). *Reasoning and Choice: An Epistemic Course in Game Theory*. Cambridge University Press.

Perry, J. (1975a). *Personal Identity*. Berkeley: University of California Press.

Perry, J. (1975b). Personal identity, memory, and the problem of circularity. In J. Perry, editor, *Personal Identity*, pages 135–158. University of California Press, Berkeley.

Perry, J. (1975c). The problem of personal identity. In J. Perry, editor, *Personal Identity*, pages 3–30. University of California Press, Berkeley.

Pettit, P. (2003). Akrasia, individual and collective. In S. Stroud and C. Tappolet, editors, *Weakness of Will and Practical Irrationality*, pages 68–96. Oxford: Clarendon Press.

Phelps, E. and Pollak, R. A. (1986). On second-best national saving and game-equilibrium growth. *Review of Economic Studies*, **35**(2), 185–199.

Piccione, M. and Rubinstein, A. (1997). On the interpretation of decision problems with imperfect recall. *Games and Economic Behavior*, **20**, 3–24.

Quante, M. (2007). The social nature of personal identity. *Journal of Consciousness Studies*, **14**(5-6), 56–76.

Quinton, A. (1975). The soul. In J. Perry, editor, *Personal Identity*, pages 53–72. University of California Press, Berkeley.

Ramsey, F. P. (1928). A mathematical theory of saving. *Economic Journal*, **38**(152), 543–59.

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

Read, D. (2006). Which side are you on? The ethics of self-command. *Journal of Economic Psychology*, **27**(5), 681–693.

Reny, P. J. (1992). Rationality in extensive-form games. *Journal of Economic Perspectives*, **6**, 103–118.

Reny, P. J. (1993). Common belief and the theory of games with perfect information. *Journal of Economic Theory*, **59**, 257–274.

Rorty, A. O. (1976). *The Identities of Persons*. Berkeley: University of California Press.

Rosenthal, R. W. (1981). Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic Theory*, **25**, 92–100.

Russell, B. (1903). *Principles of mathematic*. New York: Cambridge University Press.

Samuelson, P. (1937). A note on measurement of utility. *Review of Economic Studies*, **4**, 155–61.

Samuelson, P. (1939). The rate of interest under ideal conditions. *The Quarterly Journal of Economics*, **53**(2), 286–97.

Savage, C. W. and Ehrlich, P. (1992). *Philosophical and foundational issues in measurement theory*. Lawrence Erlbaum Associates Publishers.

Savage, L. J. (1972). *The Foundations of Statistics*. Dover Publications.

Schechtman, M. (2001). Empathic access: The missing ingredient in personal identity. In R. Martin and J. Barresi, editors, *Personal Identity*, pages 238–259. Blackwell Publishing, 2003.

Schechtman, M. (2005). Experience, agency, and personal identity. *Social Philosophy and Policy*, **22**(02), 1–24.

Schelling, T. C. (1980). The intimate contest for self-command. *The Public Interest*, **60**, 94–118.

Schelling, T. C. (1984). Self-command in practice, in policy, and in a theory of rational choice. *American Economic Review*, **74**, 1–11.

Scholten, M. and Read, D. (2006). Discounting by intervals: A generalized model of intertemporal choice. *Management Science*, **52**, 1424–1436.

Selten, R. (1975). Reexamination of the perfectness concept of equilibrium in extensive games. *International Journal of Game Theory*, **4**, 25–55.

Selten, R. (1978). The chain store paradox. *Theory and Decision*, **9**, 121–159.

Sen, A. K. (1977). Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy and Public Affairs*, **6**(4), 317–344.

Shoemaker, D. (2008). Personal identity and ethics. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, fall 2008 edition.

Shoemaker, S. (1959). Personal identity and memory. In J. Perry, editor, *Personal Identity*, pages 119–134. University of California Press, Berkeley, 1975.

Shoemaker, S. (1963). *Self-Knowledge and Self-Identity*. Ithaca: Cornell University Press.

Shoemaker, S. and Swinburne, R. (1984). *Personal Identity*. Oxford: Basil Blackwell.

Sider, T. (2000). Recent work on identity over time. *Philosophical Books*, **41**, 81–9.

Sider, T. (2001). *Four Dimensionalism*. Oxford University Press.

Sidgwick, H. (1907). *Methods of Ethics*. 7th Ed. Macmillan.

Siniscalchi, M. (2008). Epistemic game theory: Beliefs and types. In S. N. Durlauf and L. E. B. (Eds.), editors, *The New Palgrave Dictionary of Economics*. Palgrave Macmillan.

Snowdon, P. (1990). Persons, animals, and ourselves. In C. Gill, editor, *The Person and the Human Mind*. Oxford: Clarendon Press.

Stalnaker, R. (1998). Belief revision in games: Forward and backward induction. *Mathematical Social Sciences*, **36**, 31–56.

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, **38**, 332–382.

Steele, K. (2007). *Precautionary Decision-Making. An Examination of Bayesian Decision Norms in the Dynamic Choice Context*. Ph.D. thesis, University of Sydney.

Stigler, G. J. and Becker, G. S. (1977). De gustibus non est disputandum. *The American Economic Review*, **67**(2), 76–90.

Strotz, R. (1956). Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies*, **13**, 165–180.

Suppes, P. (2002). *Representation and Invariance of Scientific Structures*. Stanford: CSLI Publications.

Suppes, P. and Winet, M. (1955). An axiomatization of utility based on the notion of utility differences. *Journal of Management Science*, **1**(3), 259–70.

Suppes, P., Krantz, D. H., Luce, R. D., and Tversky, A. (1971). *Foundations of Measurement Volume II: Geometrical, Threshold, and Probabilistic Representations*. Mineola: Dover Publications.

Swistak, P. (1990). Paradigms of measurement. *Theory and Decision*, **29**(1), 1–17.

Tan, T. C. and Werlang, S. S. C. (1988). The Bayesian foundation of solution concepts of games. *Journal of Economic Theory*, **45**, 370–391.

Taylor, C. (1989). *Sources of the Self: The Making of Modern Identity*. Cambridge, MA: Harvard University Press.

Teller, P. (1975). Shimony's argument for tempered personalism. In G. Maxwell and R. Anderson, editors, *Induction, Probability, and Confirmation*, pages 166–203. University of Minnesota Press.

Thaler, R. H. and Shefrin, H. M. (1981). An economic theory of self-control. *Journal of Political Economy*, **89**(2), 392–406.

Thompson, J. J. (1997). People and their bodies. In J. Dancy, editor, *Reading Parfit*. Oxford: Blackwell Publishers.

Uzgalis, W. (2009). John Locke. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, fall 2009 edition.

von Neumann, J. and Morgenstern, O. (1944). *Theory Of Games And Economic Behavior*. Princeton University Press.

Weitzman, M. (2001). Gamma discounting. *American Economic Review*, **91**(1), 260–71.

Wilkes, K. W. (1988). *Real People - Identities without Thought Experiments*. Oxford: Clarendon.

Williams, B. (1956). Personal identity and individuation. In B. Williams, editor, *Problems of the Self*. Cambridge University Press, 1973.

Williams, B. (1970). The self and the future. *Philosophical Review*, **79**, 161–80.

Xue, L. (2008). The bargaining within. *Economics Letters*, **101**, 145–147.

Zellner, A. (1982). Reply to a comment on "Is Jeffreys a 'Necessarist'?". *The American Statistician*, **36**(4), 392–93.